Special Issue

Special Issue on
Extended Papers from Science and Information Conference 2014

SAI

www.ijacsa.thesai.org

IJACSA

WHERE WISDOM SHARES

INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS

THE SCIENCE AND INFORMATION ORGANIZATION

OAIster   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 Beta   EBSCO HOST Research Databases

Editorial Preface

The Science and Information (SAI) Organization (www.thesai.org) is an international professional organisation dedicated to promoting research, technology and development by providing multiple platforms for collaboration of professionals and researchers to share existing and generate new knowledge.

Since its inception, SAI has published thousands of articles in its journals, International Journal of Advanced Computer Science and Applications and International Journal of Advanced Research in Artificial Intelligence. Its journals are indexed by both university libraries and major international indexes. Researchers publishing papers and articles in one of SAI's journals help strengthen the support for research and technical enterprise through their work. At the same time these open access articles and papers provide a scientific voice on societal issues that are prominent today.

Each year, SAI is also responsible for organizing its annual conference, with the support of IEEE and Springer. The conference brings together some of the most devoted minds in the industry to share ideas and showcase achievements. These conferences are a place for researchers to promote and defend their work based on its merits.

SAI Conference 2014 was an overwhelming success, attracting 190+ delegates, speakers and sponsors from 50 countries and provided great intellectual and social interaction for the participants. The inspiring keynote speeches and the state-of-the-art lectures have deeply motivated attendees and envisioned future research directions.

In this special issue, we have invited thirteen of the main contributors to this conference to expand on their presentations and represent their work as an extended paper. Selected papers were invited to extend and submit them for a complete peer review process for consideration. The final decision for the inclusion in the special issue has been strictly based on the outcome of the review process. The objective of the special issue is to make available recent results and report in-progress research in the field of computer science and information technology. We hope the reader interested in the field will find this book stimulating, leading to even more contributions to the next conference.

The selected papers touch different important topics of Cloud Computing, Operating Systems, User Experience, Sentiment Analysis, Knowledge Management, Security, Business Process Management, Quantum Computing, e-Commerce, e-Learning and Simulation.

The preparation of this Special Issue has been an interesting experience. We would like to thank the authors for their patience with the process and the reviewers for providing critical evaluations of these manuscripts.

Thank you for Sharing Wisdom!

Editorial Board

Reviewer Board Members

- **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 l. 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

(iv)

- **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 Ezziyyani**
  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 Venkteshwar 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-Dubaee**
  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 SENTHILKUMAR**
  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

(vii)

CONTENTS

(ix)

An Enhanced Version of the MCACC to Augment the Computing Capabilities of Mobile Devices Using Cloud Computing

Mostafa A. Elgendy
Computer Science
Faculty of Computers and Informatics
Benha, Egypt

Ahmed Shawish
Scientific Computing
Ain Shams University
Cairo, Egypt

Mahmoud I. Moussa
Computer Science
Faculty of Computers and Informatics
Benha, Egypt

Abstract—Recently as smartphones have a wide range of capabilities a lot of heavy applications like gaming, video editing, and face recognition are now available. However, this kind of applications need intensive computational power, memory, and battery. A lot of researches solve this problem by offloading applications to run on the Cloud due to its intensive storage and computation resources. Later, some techniques chooses to offload part of the applications while leaving the rest to be processed on the smartphone based on one or two metrics like power and CPU consumption only without any consideration to other important metrics. Our previously proposed MCACC framework has introduced a new generation of offloading frameworks that handle this problem by smartly emerging a group of real-time metrics like total execution time, energy consumption, remaining battery, memory, and security into the offloading decision. In this paper, we introduce an enhanced version of the MCACC framework that can now smartly operate under low bandwidth network scenario in addition to its existing capabilities. In this framework, any mobile application is divided into a group of services, and then each of them is either executed locally on the mobile or remotely on the Cloud through a dynamic offloading decision model. The extensive simulation studies show that both heavy and light applications can benefit from the proposed framework while saving energy and improving performance compare to previous counterparts. The enhanced MCACC turns the smartphones to be smarter as the offloading decision is taken without any user interference.

Keywords—smartphones; android; offloading; mobile Cloud computing; battery; security

I. INTRODUCTION

Recently smartphones are becoming popular. Studies showed that more than 56% of users in the world use smartphone [5]. More than 53% percent of smartphone owners used the Android OS [6]. Smartphones have a wide range of capabilities like, Wi-Fi, cameras, storage, GPS and speed processors. As a result, developers are building more complex mobile applications such as into heavy applications such as natural language translators, speech recognizers, optical character recognizers, image processors and search, online games, video processing and editing, navigation, face recognition and augmented reality.

As applications become more complex, it consumes most of the mobile devices resources such as battery, memory, and computational power. Mobile applications can augment their capabilities with unlimited computing power and storage space by offloading some services to run on the Cloud; as result saving time and computation power which are called *mobile Cloud computing* [1 - 3].

A mobile Cloud computing survey show a lot of work done in this field. Some solutions considered an application as a single unit that cannot be decomposed into multiple methods and must be run either on the Cloud or locally [9], while others like [7] always offload services to execute on the Cloud all the time without taking any decisions as in. Some other solutions use a simple offloading model which take parameters like power and CPU consumption in their offloading decision as in [8-10]. Later, our proposed Mobile Capabilities Augmentation using Cloud Computing (MCACC) framework [11] has introduced a new generation of offloading frameworks that handle this problem by smartly emerging a group of real-time metrics like total execution time, energy consumption, remaining battery, memory, and security into the offloading decision. Its extensive simulation studies showed its capability to handle heavy applications by efficiently utilize the available smartphone resources and offload only when necessary based on realistic decision metrics.

In this paper, we introduce an enhanced version of the MCACC framework that can now smartly operate under low bandwidth network scenario in addition to its existing capabilities. In this framework, any mobile application is divided into a group of services, and then each of them is either executed locally on the mobile or remotely on the Cloud based on a dynamic offloading decision model. In case of low bandwidth scenario, the offloading decision is taken based on real-time comparisons between being executed locally, or compressed and then offloaded, or offloaded directly without compression.

The extensive simulation studies show that both heavy and light applications can benefit from the proposed framework even under low bandwidth scenario, while saving energy and improving performance compare to its previous counterparts.

Now, Android developers can use our proposed MCACC very easily by adding the MCACC library into their projects and by adding the MCACC builders to the project building process.

In this paper, we provide a detailed description of the whole MCACC framework with its new enhancement as a complete solution. The rest of the paper organized as follows. Section II introduces the background and shows related work. Section III describes the MCACC framework. Section IV discusses the results of the extensive simulation studies. The paper is finally concluded and future work is presented in Section V.

II. BACKGROUD

This section provides a comprehensive background on the mobile environment and application development process. In addition, it provides to a complete review on the related work done in the offloading context.

A. Mobile environment and Application Development

1) Android Architecture: *The Android platform is a software stack that was designed primarily but not exclusively to support mobile devices such as phones and tablets. This stack has several layers going all the way from low level operating system services that manage the device itself up to sample applications, things like the phone dialer, the context database, and a web browser. At the bottom, there is a i) Linux kernel layer- which provides two core services that any Android computing device will rely on. The first service is Generic operating system services which contains device drivers, memory management, process management and security. The second service is Android specific components which contains power management, Android shared memory and Inter Process Communication (IPC). Above that, there are ii) System libraries- These libraries are typically written in C and C++ and for that reason they are often referred to as the native libraries. These native libraries handle a lot of the core, performance sensitive activities on your device like quickly rendering web pages and updating the display. Beside this there is iii) Android runtime system – contains two components which support writing and running Android applications. The first component is a core Java libraries that provides a number of reusable Java building block to allow developer to write Android applications using Java programming language. The second component is the Dalvik Virtual Machine that actually executes Android applications. Above that, there's a rich iv) Application framework layer- this exposes the various capabilities of the Android OS for application developers so that they can use these capabilities in their applications. These capabilities are like package manager, window manager, view system, resource manager, activity manager, Content providers, location manager, and notification Manager. Finally at the very top, there is v) Applications - Android comes with some built-in applications which include things like the Home Screen, the Phone Dialer, the Web Browser, an Email Reader, and more. One of the* *things that are really nice about Android is that none of these apps is hardcoded into the system [12].*

2) Android Application Components: *In any android applications the following components make the structure of it, and these component are Activities, Services, Content Providers, and Broadcast Receivers, which have their own specific lifecycle within the system [12]. This study focused on activities and services as the separation between the them form a natural basis for MCACC framework.*

3) Android IPC: *When user launch an android application the operating system starts an activity that presents a graphical user interface to the user. When this activity is bound to the running service, it communicates with the service through IPC, using a predefined interface by the programmer called AIDL file and a stub/proxy pair generated by the Android pre-compiler [13]. When an activity tries to call service method, it uses the proxy object to communicate with the stub which has the actual implementation of the service as shown in Fig. 1.*

4) Android Application Development: *Any android applications have to be written in the Java. When developer writes android applications and try to build it, the build process will invoke Android Resource Manager followed by Android Pre Compiler, then it invoke Java Builder, finally invoking Package Builder to build a single APK file which can be installed on any Android device [11].*

B. Related Work

A lot of researches have been done on remote execution of mobile applications services on the Cloud to increase performance and save mobile power and memory resources [19] and [20]. These researches are divided into two paths:

1) Process and VM Migration: *In this approach a full process or full VM is migrated into the Cloud for processing. There are some researches done in this approach as follows:*

CloneCloud enables unmodified mobile applications running in an application level virtual machine to seamlessly offload part of their execution from mobile devices on device clones operating in a computational Cloud[14]. When running a complete clone of the smartphone at the remote Cloud resource, there is cost of keeping the smartphone synchronized with an application clone in the Cloud; so it's better to offload only the needed services to run on the Cloud, Also in low bandwidth network data can be compressed before offloading to the Cloud to minimize data transferred over network. **ThinkAir** exploits the concept of smartphone virtualization in the Cloud and provides method-level computation offloading [15]. *ThinkAir* creates virtual machines (VMs) of a complete smartphone system on the Cloud, and provides an online method-level offloading however it lacks flexibility and control over offloaded components. Developers organize their application using Android service design patterns. Also in low bandwidth network data can be compressed before offloading to the Cloud to minimize data transferred over network.

2) Method Offloading: *Another common approach for remote execution is to partition mobile application into some*

services that executes locally on mobile or remotely on the Cloud and this is called method offloading. There are a lot of researches which have done in this approach and these will be described and their drawbacks will be discussed in the following section.
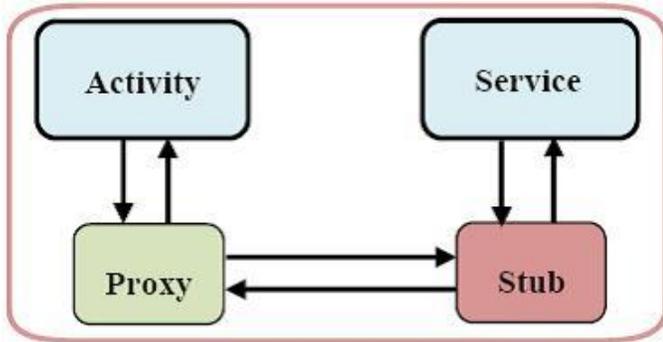
Fig.1.   An overview of the Android IPC mechanism

Cuckoo proposed a framework that automatically offloads heavy services to execute on the Cloud. *Cuckoo* use the very simple model which always prefers remote execution [7]. It's better to use some metrics in taking offloading decision like service processing time instead of offloading all the time than always offloading service directly. Also, in low bandwidth networks the time for communicating and transmit data on network and execute service on the Cloud is larger than the time to executing services on mobile, so *Cuckoo* decide to run services locally on the mobile. *Cuckoo* can compress data before sending to the Cloud, as result saving time in low bandwidth networks. Another thing, data sent to Cloud must be protected, so any security technique should be applied for protection. **Eric Chen** implements a framework that automatically offloads heavy tasks to execute on the Cloud [8]. *Eric Chen* uses total response time, energy consumption and remaining battery life in deciding whether a task should be offloaded or not without adding any memory usage consideration and security to the offloading model. Also in low bandwidth networks, application can also offload data to the Cloud by compressing data before offloading which lead to minimize data transferred over network. **Vinod** proposed a model for deciding whether to offloads heavy backend tasks to execute on the Cloud [9]. *Vinod* take some considerations like memory usage consideration and security to the offloading model. *Vinod* considered an application as a single unit that cannot be decomposed into multiple methods and must be run either on the Cloud or locally. However in some cases it's better to offload some methods to execute on the Cloud and run the others on mobile. **Karthik Kumar** provides simple analysis for deciding whether to offload computation to a server or not. This analysis tries to measure the power of sending computations to the Cloud and the power of executing computation on mobile device [1]. Although this analysis solved the problem, it lack any memory usage consideration and battery consideration when making analysis. *Kumar* also conclude that offloading data intensive tasks to the Cloud depends on the network bandwidth as if the network is low, it will better to execute service locally on the mobile and if the network is high, it will better to execute service remotely on

the Cloud. However in low bandwidth networks application may get rid of Cloud by compressing data before offloading, as result execution time and power consumption can be save **Kiran I. Koshy** try to measure energy benefits of offloading tasks from mobile devices to powerful remote servers. *Kiran* measured the energy consumed by mobile and added network energy consumed to it, and measure the energy consumed by Cloud and compared for deciding whether a task offloading reduced energy or not [10]. *Kiran* missed some metrics when making this investigation like memory usage. Another thing, data sent to Cloud must be protected, so any security technique should be applied for protection. *Kiran* can be improved by compressing data before offloading to the Cloud in low bandwidth networks. **Phone2Cloud** [21] use a naive history based method to predict average execution time of an application on smartphone. It monitor network bandwidth and leverages average CPU workload got from the resource monitor and input size of the application to predict execution time using the history log. However in data intensive application and low bandwidth network, *Phone2Cloud* always prefer to run service locally on the mobile. *Phone2Cloud* can improve his framework by compressing data before offloading to the Cloud in low bandwidth networks.

III.   PROPOSED FRAMEWORK DESIGN

In this section, the enhanced MCACC architecture is addressed in detail with its dynamic offloading model. The process done on the Cloud side and the communication between the mobile and the cloud is also discussed. Finally, this section describes the builders added to allow any android application to make use of our framework.

A. MCACC Architecture

As shown in Fig. 2. MCACC consists of four main components *i) Decision Manager - ii) Offloading Manager - iii) Execution profile - and iv) Cloud Manager*. The first three components are deployed on the mobile and the *Cloud manager* component is deployed on the Cloud. In order to use MCACC, the application should be structured using android services pattern. Note that Communication between activities and services done through stub/proxy generated by Android pre-compiler.

1) Offloading Manager: is responsible for executing the application services based on the decision taken by Decision Manager. If the decision is to execute the service locally on the mobile, then Offloading Manager calls the local service implementation from the mobile side. However if the decision is offloading the service for execution on the Cloud, then the Offloading Manager connect to the Cloud Manager and send any data needed to execute the service, Then it waits until the Cloud Manager execute the service on the Cloud and send the result back to the mobile side. At the end Offloading Manager is responsible for receiving the returned results and delivering it to the application.

2) Execution Profiler:is a profile created for each service by Decision Manager at the first of its run to store some data related to each service like execution time, power comsumption and memory consumption. It store these data for

each service based on executing sample example of the service in the following three scenarios. The first scenario is when service executed locally on the mobile. The second scenario is when service offloaded and executed on the Cloud. Note in this case it store only Cloud execution time removing the time to send data over network, as the time of sending data over network depends on network bandwidth, so it can be calculated when trying to call the service. The third scenario is when compressing service data, then offloading it for execution on the Cloud. In subsequent runs of the Decision Manager, data stored in Execution Profiler and network bandwidth will be used in taking the offloading decision.
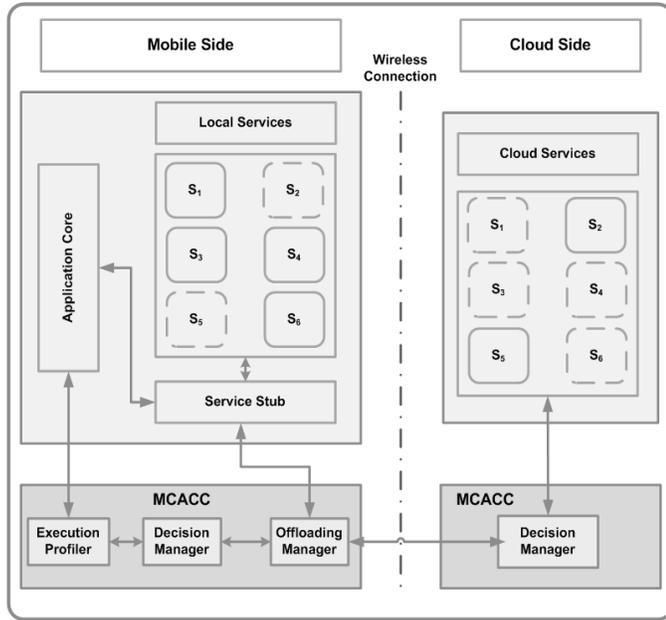
Fig.2. MCACC Architecture – services like S2 can be offloaded to run on the Cloud using Offloading Manager and Cloud Manager.

3) Decision Manager: uses a dynamic offloading model to decide at runtime weather the service will be offloaded to the Cloud – offloaded with or without compression, or executed locally on the mobile. First it get the network bandwidth, then it read service stored data about execution time, power comsumption and memory consumption from the execution profiler for the three running scenarios. Finally it uses the offloading model algorithm described in Section B to make an offloading decision. When it decides to run the application locally or remotely, it calls Offloading Manager which is reasonable for service execution.

4) Cloud Manager: is reasonable for service execution on the Cloud. In the first run it receives the Jar file which contains the remote implementation of all application services and the needed libraries from Offloading Manager and install it at the Cloud side. At any time when the Offloading Manager try to call service from the Cloud, Cloud Manager receive all required data to execute the service, execute it and return the result to Offloading Manager.

B. Offloading Model

When an activity invokes a method of a service, the Android IPC mechanism directs this call through the proxy and the kernel to the stub. In normal android application the stub invokes the local implementation of the method and then returns the result to the proxy. When using the MCACC, the android application becomes smarter. MCACC uses dynamic offloading model to evaluate whether it is beneficial to offload the method to run on the Cloud, compressing it and then offloading to run on the Cloud, or executing it locally on mobile. MCACC uses five metrics in taking decision. These metrics are *i)* execution time- *ii)* energy consumption- *iii)* remaining battery life - *iv)* memory usage and *v)* security.

The **execution time metric**, dealt with total time required to perform a task. Let I be the number of instructions involved in a method invocation, S_{Mobile} be the processor speed (instructions per second) of the mobile and S_{Cloud} be the processor speed on the Cloud. If the amount of data transferred between mobile and Cloud is D and the network bandwidth is B, the time it takes to transfer data is D/B. Using these, MCACC derived the relationship between execution speed and communication overhead as shown:

Let $T_{Mobile} = \frac{I}{S_{Mobile}}$, $T_{Cloud} = \frac{I}{S_{Cloud}}$ and $T_{Net} = \frac{D}{B}$ .from these we can drive this

$$\text{If } T_{Mobile} - (T_{Net} + T_{Cloud}) > 0$$
$$T=1$$
$$\text{else If } T_{Mobile} - (T_{Net} + T_{Cloud}) > 0 \text{ after compression}$$
$$T=1$$
$$\text{else}$$
$$T=0 \tag{1}$$

If the execution time on the mobile is greater than the sum of the time to send data over network and execution time on the cloud or if this is true but in the case of compressing data before sending to the Cloud, then it's beneficial to offload to run on the Cloud. Other than theses case it's better to execute service locally on the mobile.

The **energy consumption metric**, dealt with energy consumption. Let P_{Mobile} *watt* the energy consumed by mobile for computing per second, P_{Cloud} *watt* the energy consumed by mobile for being idle until executing service on the cloud per second and P_{Net} *watt* for sending and receiving data; then the energy consumed is $P_{MTot} = P_{Mobile} \times T_{Mobile}$ *watt*. If the Cloud performs the computation, the energy consumed was $P_{NTot} = P_{Net} \times T_{Net}$ *watt* for the communication overhead and $P_{CTot} = P_{Cloud} \times T_{Cloud}$ *watt*. Using these MCACC derived the relationship between energy consumption on the mobile and on the Cloud:

$$\text{If } P_{MTot} - (P_{NTot} + P_{CTot}) > 0$$
$$P = 1$$
$$\text{else if } P_{MTot} - (P_{NTot} + P_{CTot}) > 0 \text{ after compression}$$
$$P = 1$$
$$\text{else}$$
$$p = 0 \tag{2}$$

If the energy consumption on the mobile is greater than the sum of the energy consumed to send data over network and execution time on the cloud or if this is true but in the case of compressing data before sending to the Cloud, then it's beneficial to offload to run on the Cloud. Other than theses case it's better to execute service locally on the mobile.

The ***remaining battery metric***, dealt with mobile remaining battery life in decision-making. Let L be the mobile remaining battery life in *watt*. If a task couldn't be completed with the remaining battery of the mobile or if the remaining battery is sufficient to upload the input data required to perform the task on the Cloud, the Cloud can do the task while the mobile's battery drains out. The Cloud can later return the results to mobile. This condition expressed by the following metric:

$$If \quad L - (P_{NTot} + P_{CTot}) > 0$$
$$B = 1$$
$$else\ if\ L - (P_{NTot} + P_{CTot}) > 0\ after\ compression$$
$$B = 1$$
$$else$$
$$B = 0 \tag{3}$$

$$If \quad P_{MTot} - L > 0$$
$$B = 1$$
$$else$$
$$B = B \quad - from\ previous\ equation \tag{4}$$

The ***memory usage metric***, dealt with memory used to perform a task. Let mem_{avail} the memory available on mobile, mem_{total} the total memory available and mem_{th} is the percentage of threshold that the process will not exceed. If the service memory usage exceeded the threshold specified with or without compression, the service offloaded for execution on the Cloud and return the results to mobile, otherwise the service will be executed locally on the mobile. This condition expressed by the following metric:

$$mem_{calc} = \frac{mem_{avail}}{mem_{total}} \times 100$$

From the above equation

$$If \quad mem_{th} - mem_{calc} > 0$$
$$M = 1$$
$$else\ if\ mem_{th} - mem_{calc} > 0 \quad after\ compression$$
$$M = 1$$
$$else$$
$$M = 0 \tag{5}$$

The ***security metric*** that MCACC use when we try to offload dealt with security used; is the user needs a security on data before sending to the Cloud, using this metric allows to encrypt data before sending and decrypt it on the Cloud for processing. MCACC uses ***AES*** technique for encryption and decryption.

$$If\ user\ need\ security$$
$$S = 0$$

$$else$$
$$S = 1 \tag{6}$$

After calculating *T*, *P*, *B*, *M* and *S* from these previous metrics mobile user can set priorities to each metric using the following weights w_t, w_P, w_l, and w_m, finally the offloading model decide whether the service will be offloaded to the Cloud – offloaded with or without compression, or executed locally on the mobile using the following equation:

$$Let \quad C = T \times w_t + P \times w_p + B \times w_l + M \times w_m$$
$$If\ C > 0.5$$
$$Cloud = 1.$$
$$else$$
$$Cloud = 0. \tag{7}$$

After calculating *C* from the previous equation and selecting whether needing security or not. If *C* is greater than 0.5, then the service will be executed on the Cloud, otherwise the services will be executed locally on the mobile. When offloading the service for execution on the Cloud if the user select security, then data will be encrypted before sending it to the Cloud.

C. Cloud Side

Cloud Manager is written with pure Java so any application can offload its computation to any resource running a Java Virtual machine; either being machines in a commercial Cloud such as Amazon EC2 [17] or private Clouds such as laptops and desktops. MCACC run *Cloud Manager* which handles all offloading requests from the clients, installation of offloaded services and their initialization, libraries needed and. Finally *Cloud Manager* invokes services when *Offloading Manager* needs to call them. Note that at first run of user application MCACC sends the jar file created by *Jar Creator* to the Cloud.so all mobile services become available for execution on the Cloud.

D. Communication: IBIS

In order to execute methods on a remote resource, the phone has to communicate with the Cloud resource. MCACC used the Ibis communication middleware for this purpose [11]. The Ibis middleware consists of two subsystems, the Ibis Distributed Deployment System and the Ibis High-Performance Programming System. MCACC framework has been implemented on top of the Ibis High Performance Programming System, which offered an interface for distributed applications [16].

E. Integration into Build Process

In any Android application the connection between the activities and AIDL services processed as follow: When an activity needs to invoke a method in a service, it makes call to the matching method in the proxy. The proxy is responsible for connecting to service to call the need method. The proxy doesn't connect to service directly but, it connects to stub which call the local service and return the result to the proxy. The proxy takes this result and passes it to the caller activity. The framework is deployed in the application layer without modifying the underlying Android platform. The framework

provided three Eclipse builders that can be inserted into an Android project's build configuration in Eclipse.

1) Stub Modifier: The first builder is called the Stub Modifier and has to be invoked after the Android Pre Compiler, but before the Java Builder. The Stub Modifier will rewrite the generated Stub for each AIDL interface, so that at runtime it connected to the Decision Manager to take offloading decision whether a method will be invoked locally on mobile or remotely on the Cloud.

2) Remote Creator: The second builder called Remote Creator used to derive a dummy remote implementation from the available AIDL interface for each service. Now the application with two copies of a service during the build process: i) the first copy of the service added by Android called the local service that executes on the mobile. -ii) the second copy of the service added by framework using Remote Creator and contains the same implementation as the local services and called remote service. This second copy will be executed on the Cloud, so developer can change its implementation to use all Cloud resources like parallel processing.

3) Jar Creator: The third builder called Jar Creator used to build a Java Archive File (jar) which contains the remote implementation and all needed libraries. This jar file will be installed on the Cloud. The Remote Creator and the Jar Creator have to be invoked after the Java Builder, but before the Package Builder, so that the jar will be part of the Android Package file that results from the build process as shown in Fig. 3.

IV. SIMULATION STUDIES

To evaluate the MCACC framework, a face detection application was used. It is an application that allow user to select image from gallery or to take real-time one, then the application execute face detection service locally on mobile or remotely on the cloud using the proposed enhanced MCACC framework. After that detection service return an array of all detected faces. Finally the application use this array to draws a rectangle around each detected face as shown in Fig.4. This application uses *JavaCV* library to detect image faces. *JavaCV* is a wrapper that allows accessing the *OpenCV* library directly from within Java Virtual Machine (JVM) and Android platform.

A. Simulation Setup

Hardware: On the mobile side a Samsung Galaxy S Advance GT-I9070 mobile was used. The mobile uses Android operating system in version 4.1.2, integrates with Wi-Fi interface, and a battery capacity of 1500mAh. It has CPU with 1 GHz, 1.97 GB system storage and 3.92 GB USB storage at 3.7 volts. On the Cloud side a laptop with a core I3 2.13 processor, 4 Giga Ram acted as a Cloud provider. We evaluate the execution time, power consumption and CPU consumption for our application. To measure the power consumption, CPU consumption, and used memory a software called little eye V2.4.0.0 is used [18].
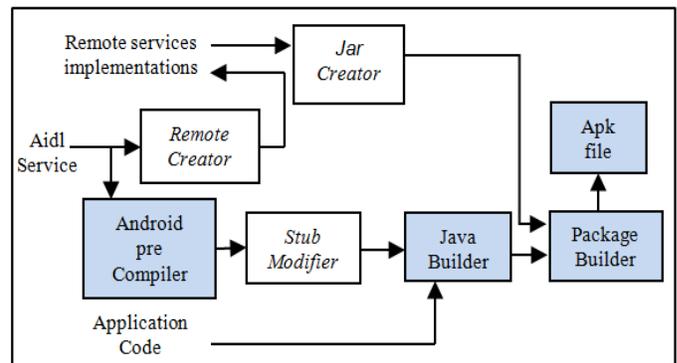
Fig.3. A schematic overview of how components integrate into the default build process.

B. Result and Discussion

Five images were used in the evaluation of the face detection application. The application was evaluated three times. *i)* First the application was evaluated in a good bandwidth network under two scenarios; the first one represents the execution of the face detection service on the mobile device, while the second one represents the offloading of the service for execution on the Cloud. *ii)* Second the application was evaluated in a low bandwidth network under three scenarios; the first one represents the execution of the face detection service on the mobile device, the second represents the offloading on the Cloud and the third represents compressing the data before offloading it on the Cloud. and finally *iii)* the application was evaluated using more than on security algorithms under two scenarios; the first one represents the execution on the mobile device, while the second one represents the offloading on the Cloud.
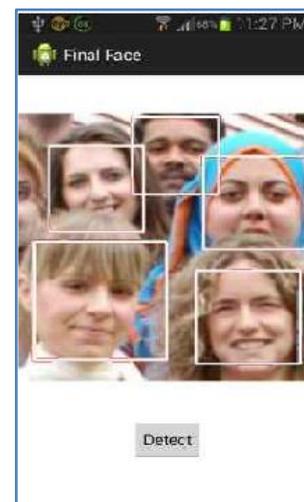
Fig.4. Screenshot of face detection application

1) Good bandwidth network: Five images were used in evaluating the face detection application under two scenarios; the first one represents the execution of the face detection service on the mobile device, while the second one represents the offloading of the service for execution on the Cloud.

Fig. 5 shows the execution time where the x -axis represents the size of the images in kilo bytes while the y-axis represents the processing time in seconds. It can be easily noted that the increase of image's size implies a corresponding increases in the processing time on the mobile device consuming memory and power resources while in the offload scenario (i.e., using the Cloud) such resources are relatively preserved.

For example, the image with size 9830.4 kb takes about 7 seconds to be executed on mobile while it takes about 5 second to be executed on Cloud. However, it is worth to note that the time in the offloading scenario Cloud is the sum of the time needed to send/receive the service to/from the Cloud plus the execution time there. So the offloading scenario does not only depends on the Cloud execution time but also depends on the network bandwidth.

Fig.6. shows the CPU consumption percentage in both scenarios. The x -axis represents the size of images in kilo bytes and the y-axis represents the average of CPU consumption percentage. The result demonstrates the aggressive consumption of the mobile resources in case of executing such heavy service. It also shows the efficiency of the offloading approach to save such resources. For example, the execution of face detection service on mobile consumed about 33% of CPU, while this percentage is minimized to 7.5% in the offloading scenario.

Fig. 7 and 8. describe the power and memory consumption in both scenarios, respectively. The x -axis represents the size of images in kilo bytes and y-axis represents the power and memory consumed by the mobile. The results of both experiments match well with the conclusion of the previous one: offloading is a better choice in case of heavy services. However, we are not arguing to prove this conclusion, we are here providing a smart offloading framework that is able to take the right decision under any circumstance, taking into consideration all of the above real-time metrics.
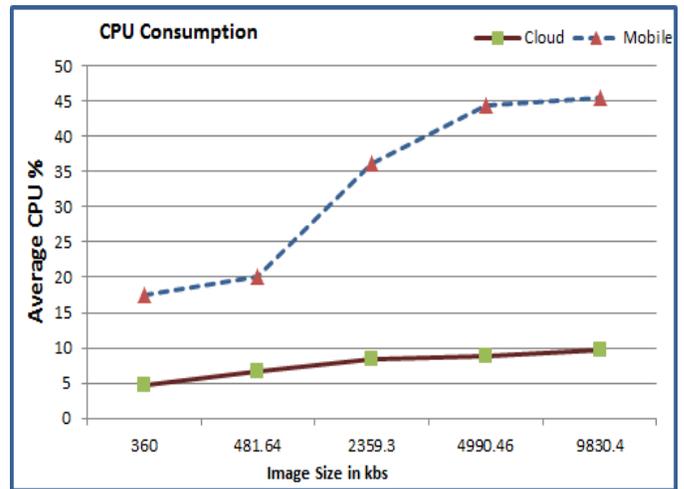
Fig.6.   CPU consumption percentage on mobile and on Cloud
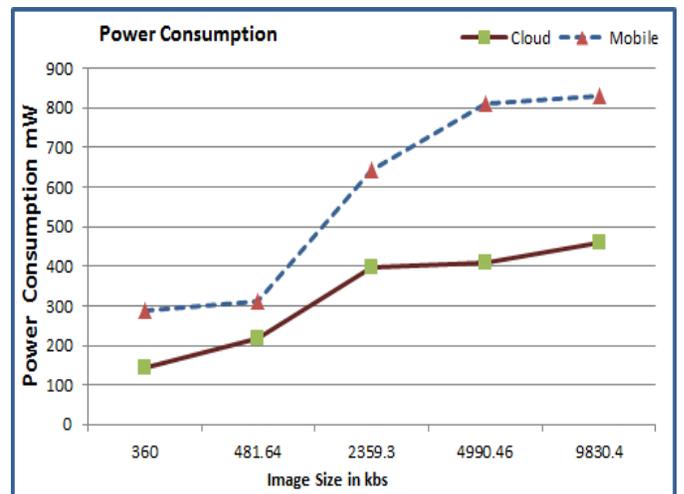
Fig.7.   Power consumption on mobile and on Cloud
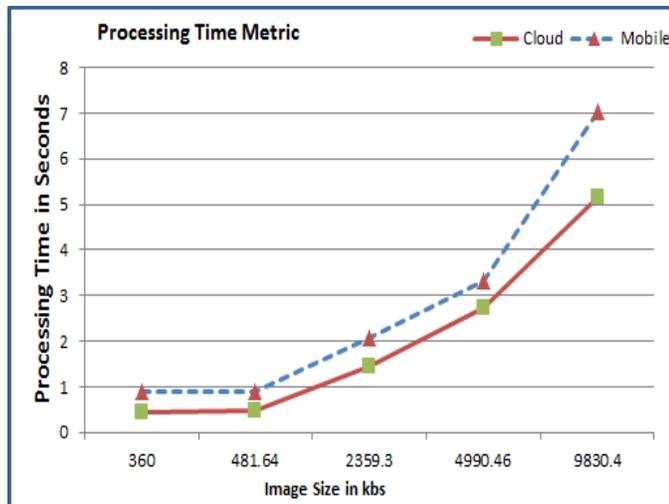
Fig.5.   Processing time on mobile and on Cloud
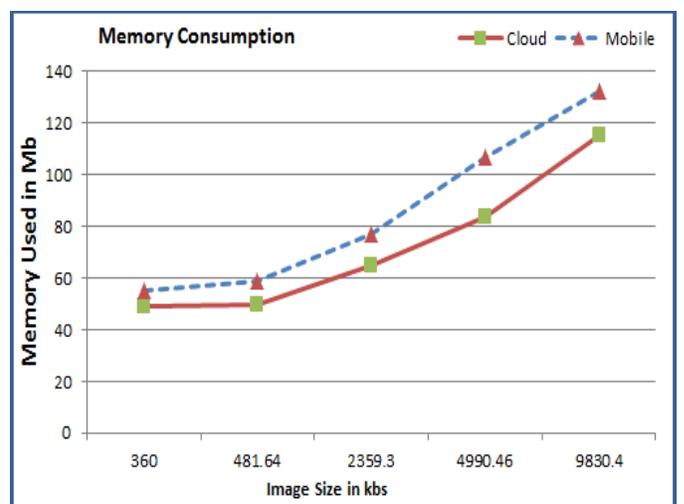
Fig.8.   Memory consumption on mobile and on Cloud

1) Low bandwidth network: In this part the application is evaluated in a low bandwidth network under three scenarios; the first one represents the execution of the face detection service on the mobile device, the second represents the offloading on the Cloud, and the third represents compressing the data before offloading.

Fig.9. shows the execution time in the three scenarios in low bandwidth networks. The x -axis represents the size of the images in kilo bytes and the y-axis represents the processing time in seconds. It can be easily noted that the execution time on the Cloud without compression is greater than execution time on mobile and on the Cloud with compression as it takes more time to transfer data through the low bandwidth network.

For example, the image with size 9830.4 kb takes about 12 seconds to be executed on Cloud while it takes about 7 second when executed on the mobile. Similarly, as the image size increases, the execution time on the Cloud without compression increases too in comparison with the other two scenarios. It is also noted that the execution time on the mobile is nearly equal to the compression scenario. For example, the image with size 9830.4 kb takes about 7 second to be executed on the mobile and almost the same when offloaded on the Cloud with compression. Accordingly, we conclude that compressing data and offloading it will give the same performance as processing the requested service on the mobile; nevertheless it will save the mobile resources.

Fig.10. shows the CPU consumption percentage in the three scenarios. The results demonstrate the aggressive consumption of the mobile resources in case of executing such heavy service locally on the mobile. It also shows the efficiency of offloading service to save such resources.

For example an image with size 9830.4 kb consumes about 48% of the mobile CPU in the first scenario while consuming 10% and 16% in the second and third scenarios, respectively. It also noted that the execution of face detection service on mobile consumed about 34% of CPU on average, while this percentage is minimized to 12.45% in the compression offloading scenario and 7.4 % in the offloading scenario without compression. Accordingly, it can be concluded that in low bandwidth networks, if the user priority is to save the mobile CPU consumption, then it is better to offload service to the Cloud with or without compression.

Fig.11. describes the power consumption in the three scenarios, respectively. The results match well with the conclusion of the previous one; offloading data to the Cloud or compressing data and then offloading to the Cloud is a better choice in case of heavy services if the network bandwidth is low. the extensive simulation studies report that in low bandwidth network it is better to compress the data before offloading to the Cloud.

2) Security: the application was evaluated with more than on security algorithms using two scenarios; the first one represents the execution on the mobile device, while the second one represents the offloading on the Cloud.
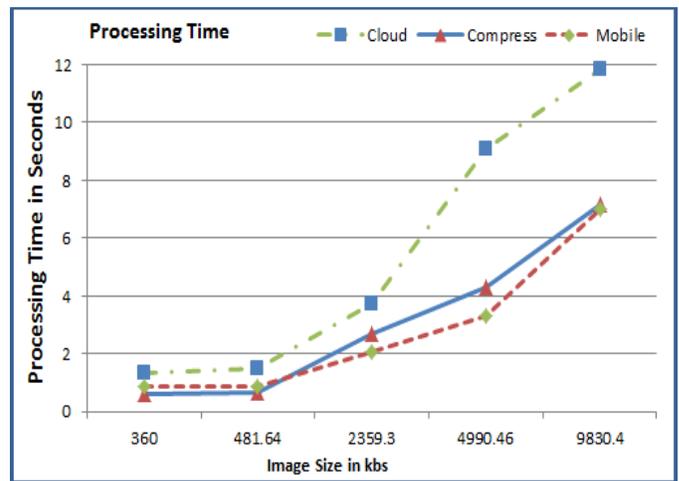

Fig.9. Processing Time of the application under the three scenarios using low bandwidth network
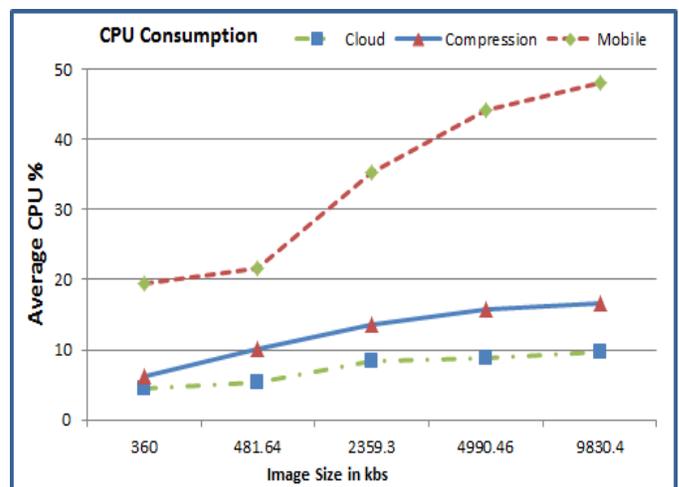
Fig.10. CPU Consumption of the application under the three scenarios using low bandwidth network.
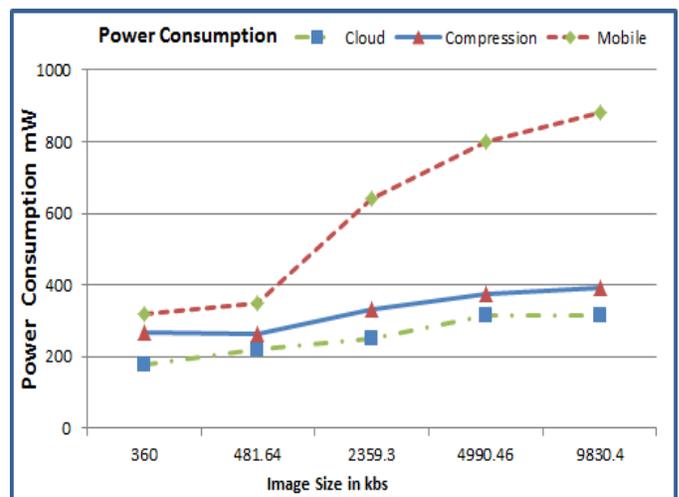
Fig.11. Power Consumption of the application under the three scenarios using low bandwidth network

Fig.12. shows the execution time when running the application on the mobile versus a secured offloading. As usual, the x -axis represents the size of images in kilo bytes and y-axis represents the processing time in millisecond. It should be noted here that before offloading the service (i.e., sending the faces image to be detect on the Cloud) the MCACC will first encrypt it and then decrypt it after receiving the results from the Cloud. The results shows that adopting a security layer on the transmitted data add an overhead on the mobile resources in order to be encrypt and decrypt. In some case this processing time is acceptable in small sized images like for example image of size 360 kb takes (0.583 second) in the secured offloading scenario while it takes (0.885 second) on the mobile. In other case like for example image of size 9830.4 kb, it takes (7.49 seconds) in the secured offloading scenario while it takes (7.012 seconds) on the mobile. It is also worth to note that the AES technique for encryption and decryption is better than the Blowfish technique as shown it the figure. We can easily note the effect of security layer on the offloading process and how it may affect the processing time on the mobile.

In general, the extensive simulation studies report that executing application services on the mobile consumed a lot of the mobile's resources which is not acceptable, while offloading it to the Cloud may save such resources. Also the results showed that in low bandwidth network, application services can be offloaded by compressing data before offloading. The results also showed that when adding a security layer to the offloading process an additional overhead should be taken into consideration. On the other hand, it is worth to note that the proposed framework supports automatic offloading of multiple Android services based on a group of realistic metrics inspected instantaneously from the smartphone. In addition, the with the popular open source Android framework and the Eclipse development tool. It provides a simple programming model, familiar to developers. This model allows developer to use our framework very easily and adds offloading components automatically.
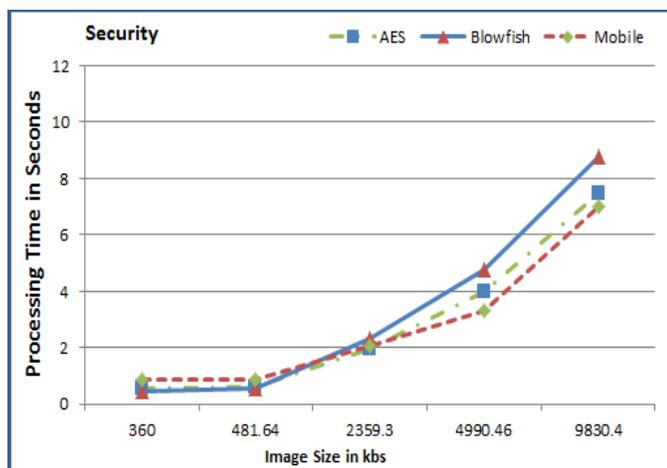
Fig.12. Processing time on mobile and on Cloud when Appling encryption techniques

The Proposed MCACC is efficiently solving a group of drawbacks in the current available techniques. For example, it overcomes the Clone Cloud [14] deficiency by offloading only the needed services based on the offload model, and hence avoids the costly process of keeping the smartphone synchronized with an application clone in the cloud. Moreover, by adopting a runtime offloading model based on five metrics, MCACC is smarter than Cuckoo [7] that uses a very simple heuristic approach to always send services to be executed on cloud without any decision. With respect to the solutions provided by Eric Chen [8], Vinod Namboodiri [9], Karthik Kumar [1], and Kiran I. Koshy [10] that utilize metrics like total response time, energy consumption and battery power in their offloading decision, MCACC is still better as it additionally utilizes the memory and security metrics. Also MCACC use cloud also in low bandwidth network by compressing data before offloading to the Cloud, so saving mobile resources.

V. CONCLUSION

This paper proposed an enhanced version of the framework called Mobile Capabilities Augmentation using Cloud Computing (MCACC) that helps smartphone to handle heavy applications. The new enhancement extends the previous framework capabilities to utilize the limited available resources of the smartphones and smartly offload the services to the Cloud even under low bandwidth scenario. In this framework, any mobile application is divided into a group of services, and then each of these services are either executed locally on the mobile or remotely on the Cloud using a dynamic offloading decision model. Here, the decision is based on real-time metrics: total execution time, energy consumption, remaining battery, memory, security, and network bandwidth.

The extensive simulation studies report the ability of the proposed framework to efficiently utilize the available smartphone's resources in addition to augmenting them using the Cloud Computing.Our future work will focus on enabling parallelization of the offloaded services and minimizing the security overhead between the mobile and the Cloud.

REFERENCES

[1] K. Kumar, and L.Yung-Hsiang, "Cloud Computing for Mobile Users: Can Offloading Computation Save Energy?," *Computer* , vol.43, no.4, pp.51-56, April 2010 doi: 10.1109/MC.2010.

[2] P. Mell, and G. Timothy, "The NIST definition of Cloud computing," *NIST special publication*, 800(145), 7. 2011.

[3] D. Kovachev, C. Yiwei, and K. Ralf, "Mobile Cloud Computing: A Comparison of Application Models, " *Computing Research Repository*, abs-1107-4940, 2011.

[4] A. Berl, G. Erol, D. G. Marco, G. Giovanni, D. M. Hermann, Q. D. Minh, and P. Kostas, "Energy-Efficient Cloud Computing," *The Computer Journal* vol. 53, pp. 1045-1051. 2010.

[5] webpage on Smartphone users. [Online]. Available: http://www.go-gulf.com/blog/smartphone/. 2012.

[6] Hepburn, A. webpage on Android usage. [Online]. Available: http://www.digitalbuzzblog.com/infographic-2013-mobile-growth-statistics/. 2013.

[7] R. Kemp, P. Nicholas, K. Thilo, and B. Henri, "Cuckoo: A Computation Offloading Framework for Smartphones," *Mobile Computing, Applications, and Services.*, vol. 76, pp. 59-79, 2012.

[8] E. Chen, O. Satoshi, and H. Keitaro, "Offloading Android applications to the Cloud without customizing Android," in *Pervasive Computing and Communications Workshops IEEE International Conference on*, pp. 788-793, 19-23 Mar 2012.

[9] V. Namboodiri, and G. Toolika, "To Cloud or not to Cloud: A mobile device perspective on energy consumption of applications," in *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2012 IEEE International Symposium on a*, pp. 1-9, 25-28 Jun 2012.

[10] K. I. Koshy, M. J. Andrew, N. Andrew, and O. Michael, "Can Cloud computing lead to increased sustainability of mobile devic?," *Sustainable Systems and Technology (ISSST), 2012 IEEE International Symposium on*, pp.1-4, 16-18 May 2012.

[11] M. A. Elgendy, S. Ahmed, and I. M. Mahmoud, "MCACC: New Approach for Augmenting the Computing Capabilities of Mobile Devices with Cloud Computing", *Science and Information Conference (SAI),* pp 79-86, 27-29 Aug 2014.

[12] M. Gargenta, Learning Android Building Applications for the Android Market, O'Reilly Media, 2011.

[13] Webpage on Android Developer AIDL [Online]. Available: http://developer.android.com/guide/components/aidl.html. 2013.

[14] B. G. Chun, I. Sunghwan, M. Petros, N. Mayur, and P. Ashwin, " CloneCloud: elastic execution between mobile device and Cloud, " *6th conference on Computer systems (EuroSys),* pp. 301–314, 2011.

[15] S. Kosta, A. Andrius, H. Pan, M. Richard, and Z. Xinwen, "ThinkAir: Dynamic resource allocation and parallel execution in the Cloud for mobile code offloading," *INFOCOM, 2012 Proceedings IEEE*, pp. 945-953, 25-30 Mar 2012.

[16] R. V. V. Nieuwpoort, M. Jason, H. Rutger, K. Thilo, and E. B. Henri, "Ibis: an Efficient Java-based Grid Programming Environment," in *Joint ACM Java Grande - ISCOPE 2002 Conference*, pp. 18-27, 2002.

[17] The Amazon Elastic Computing website. [Online]. Available: http://aws.amazon.com/ec2/. 2013.

[18] The Little eye website. [Online]. Available: http://www.littleeye.co/. 2014.

[19] E. Cuervo, B. Aruna, K. C. Dae, W. Alec, S. Stefan, C. Ranveer, and B. Paramvir, "MAUI: making smartphones last longer with code offload," in *International Conference on Mobile Systems, Applications, and Services*, pp. 49-62, 2010.

[20] M. Shiraz, G. Abdullah, H. K. Rashid, and B. Rajkumar, "A Review on Distributed Application Processing Frameworks in Smart Mobile Devices for Mobile Cloud Computing," *Communications Surveys & Tutorials, IEEE* , vol.15, no. 3, pp. 1294-1313, Third Quarter 2013.

[21] Xia, F., Ding, F., Li, J., Kong, X., Yang, L.T., Ma, J., "Phone2Cloud Exploiting computa-tion offloading for energy saving on smartphones in mobile Cloud computing", In: Information Systems Frontiers, vol.16, no. 1, pp. 95-111, (2014).

Design of a Content Addressable Memory-based Parallel Processor implementing (−1+j)-based Binary Number System

Tariq Jamil

Department of Electrical and Computer Engineering
Sultan Qaboos University
AlKhod 123, Muscat, OMAN

Abstract—**Contrary to the traditional base 2 binary number system, used in today's computers, in which a complex number is represented by two separate binary entities, one for the real part and one for the imaginary part, Complex Binary Number System (CBNS), a binary number system with base (−1+j), is used to represent a given complex number in single binary string format. In this paper, CBNS is reviewed and arithmetic algorithms for this number system are presented. The design of a CBNS-based parallel processor utilizing content-addressable memory for implementation of associative dataflow concept has been described and software-related issues have also been explained.**

Keywords—binary number; complex binary; parallel processing; content-addressable; memory; associative dataflow; compiler; operating system

I. INTRODUCTION

A complex number consists of two components, namely the *real* part and the *imaginary* part, and it represents a point in a two-dimensional space. The real part is used to plot the position of the point along the horizontal axis while the imaginary part represents the position of the same point along the vertical axis. In today's computers, a complex number is stored in base 2 binary representation with both real and imaginary parts of the number represented individually. Thus, an arithmetic operation between two complex numbers becomes the accumulation of results from two sub-arithmetic operations on each pair of real and imaginary parts of the two given complex numbers. This increases the execution time of the arithmetic operations for complex numbers and hence causes delay in generating the output in computer applications where complex numbers are frequently used, such as image processing and signal processing. Efforts to represent the whole complex number (both real and imaginary parts) as single binary string date back to 1960s when D. Knuth proposed an imaginary-base binary number system with base $2j$ [1] and W. Penney attempted to define a number system, first with base −4 and then by using a complex number $(−1 + j)$ as the base [2,3]. The main problem encountered with using these bases at that time was the inability in formulating an efficient division process and the exorbitant cost of memory to store long string of binary numbers representing complex numbers in the new bases. In 1996, V. N. Stepanenko defined a number system with base $j\sqrt{2}$ in which even powers of the base yielded real numbers and odd powers

of the base resulted in imaginary numbers [4]. Although he was able to resolve the division problem in his proposed number system as an "all-in-one" operation, yet in his algorithm, "everything…reduces to a good choice of an initial approximation…" in a Newton-Raphson iteration which may or may not converge. T. Jamil *et. al.* have revisited Penney's proposed $(−1 + j)$ -base number system and have done extensive mathematical analysis of *Complex Binary Number System* (CBNS) to establish it as a viable binary number system for representing complex numbers in the computer systems [5,6,7]. In 2001/2003, D. Blest and T. Jamil have presented an efficient division algorithm for complex binary numbers which has paved the way for implementation of this unique number system in computer hardware for all types of arithmetic operations [8,9]. This paper is intended to be a brief review of the CBNS and then it describes the architecture of an associative dataflow processor which amalgamates the parallelism inherent in content-addressable memories with the complex binary representation of complex numbers, a design which has been granted an innovation patent by the Australian Patent Office in 2010 [10].

This paper is organized as follows: In Section II, review of complex binary number system and algorithms for arithmetic operations for this number system are presented. In Section III, the concept of associative dataflow is described, which is followed by the design of *Complex Binary Associative Dataflow Processor* (CBADP) in Section IV. The software design issues related to CBADP are presented in Section V. Conclusions and suggestions for further research are outlined in Section VI. Acknowledgment and references are listed at the end of this paper.

II. REVIEW OF COMPLEX BINARY NUMBER SYSTEM (CBNS)

A. Conversion Algorithms [7]

- To represent a given position integer N in CBNS, the following steps are followed: (i) Express N in terms of power of 4 using repeated division process. (ii) Convert the base 4 number $(...q_5q_4q_3q_2q_1q_0)$ to base −4 by replacing each digit in odd location $(q_1, q_3, q_5 ...)$ with its negative to get $(...−q_5, q_4, −q_3, q_2, −q_1, q_0)$. (iii) Normalize the new number (i.e., get each digit in the range 0 to 3) by repeatedly adding 4 to the negative

digits and adding a 1 to the digit on its left. If the digit is 4, replace it by a zero and subtract a one from the digit on its left. (iv) Now replace each digit in base –4 representation with the corresponding four-bit sequence (0 →0000; 1→ 0001; 2→ 1100; 3→ 1101). To convert a negative integer into CBNS representation, we simply multiply the representation of the corresponding positive integer with 11101 (equivalent to $-1_{base\ -1+j}$) according to the multiplication algorithm given in sub-section *B*.

- To obtain CBNS representation of imaginary numbers, we simply multiply the CBNS representation of corresponding positive number with 11 (equivalent to $(+j)_{base10}$) or 111 (equivalent to $(-j)_{base10}$) according to the multiplication algorithm given in sub-section *B*.

- To represent a fraction *F* in CBNS, we first express the fraction in terms of powers of $\frac{1}{2} = 2^{-1}$ such that $F = r_0 = f_1. 2^{-1} + f_2. 2^{-2} + f_3. 2^{-3} + f_4. 2^{-4} + \dots$ to machine limit. The coefficients f_i and the remainders r_i are given as follows: Initially, if $2r_0 - 1 < 0$ then $f_1 = 0$ and set $r_1 = 2r_0$ or if $2r_0 - 1 \geq 0$ then $f_1 = 1$ and set $r_1 = 2r_0 - 1$. Then, if $2r_i - 1 < 0$ then $f_{i+1} = 0$ and $r_{i+1} = 2r_i$ or if $2r_i - 1 \geq 0$ then $f_{i+1} = 1$ and $r_{i+1} = 2r_i - 1$. We continue this process until $r_i = 0$ or the machine limit has been reached. Then, for $\forall f_i = 1$, we replace its associated 2^{-i} according to the sequence ($2^{-1} \to 1.11$, $2^{-2} \to 1.1101$, $2^{-3} \to 0.000011$, $2^{-4} \to 0.00000001$) [for $i > 4$, refer to [8]). Note that addition is according to algorithm given in sub-section *B*.

- To represent a floating-point number in CBNS, which consists of both an integer and a fraction, CBNS representation is obtained by adding the CBNS representation of each individual part according to the addition algorithm given in sub-section *B*. All rules for obtaining negative integer and positive/negative imaginary number representations, as discussed previously, are equally applicable for obtaining negative floating point and positive/negative imaginary floating point representations in CBNS.

- To represent a complex number in CBNS, we simply add the CBNS representation of real part with the CBNS representation of the imaginary part according to the addition algorithm given in sub-section *B*.

B. Arithmetic Algorithms [7]

- To add two complex numbers represented in CBNS format, the truth table is given as follows: 0 + 0 = 0; 0 + 1 = 1; 1 + 0 = 1; 1 + 1 = 1100. When two numbers with 1s in position *n* are added, this results in 1s in positions *n+3* and *n+2* and 0s in positions *n+1* and *n* in the sum. Similar to the ordinary computer rule where 1+111 … (to limit of machine) =0, we have 11 + 111 = 0 [Zero Rule] in CBNS.

- To subtract two complex binary numbers, the truth table followed is: 0 – 0 = 0 ; 0 – 1 = * ; 1 – 0 = 1 ; 1 – 1 = 0. For the case where 1 is subtracted from 0 (the * case in the rules), the following algorithm is applied: Given our

minuend is: $a_n a_{n-1} a_{n-2} \dots a_{k+4} a_{k+3} a_{k+2} a_{k+1} a_k \mathbf{0} \quad a_{k-1} \dots a_3 a_2 a_1 a_0$ and subtrahend is: $b_n b_{n-1} b_{n-2} \dots b_{k+4} b_{k+3} b_{k+2} b_{k+1} \mathbf{1} b_{k-1} \dots b_3 b_2 b_1 b_0$. Then, the result of subtracting 1 from 0 is obtained by changing: $a_k \to a_k + 1$, $a_{k+1} \to a_{k+1}$ (unchanged), $a_{k+2} \to a_{k+2} + 1$, $a_{k+3} \to a_{k+3} + 1$, $a_{k+4} \to a_{k+4} + 1$, and $b_k \to 0$.

- To multiply two complex binary numbers, we follow the same method that we use for traditional binary numbers, except that while adding intermediate summands, addition algorithm outlined previously in this sub-section is used. The Zero Rule plays an important role in speeding up the result of the multiplication operation.

- To perform the division of two complex numbers represented in CBNS, we take the reciprocal of the denominator and multiply it with the numerator as per algorithm described above. The reciprocal of the complex number is estimated using the following algorithm: Given $z = w^{-1}$, we start with our initial approximation of z by setting $z_0 = (-1 + j)^{-k}$ where k is obtained from the representation of w such that $w = \sum_{i=-\infty}^{k} a_i (-1 + j)^{-i}$ in which $a_k \equiv 1$ and $a_i \equiv 0$ for $i > k$. The successive approximations are then obtained by $z_{i+1} = z_i (2 - wz_i)$. If the values of z do not converge, we correct our initial approximation of z by setting $z_0 = j(-1 + j)^{-k}$ which will definitely converge [9].

III. ASSOCIATIVE DATAFLOW CONCEPT (ADC)

The associative dataflow concept is, in fact, an extension of the traditional concept of dataflow and is obtained by elimination of *tokens* generated during dataflow processing. In traditional dataflow, a computer program is represented as a graph consisting of nodes (representing instructions) and arcs (representing data dependencies between the nodes) [11]. The operands and their control information are conveyed from one node to another in data packets called *tokens*. The process of determining the executability of instructions is through matching of the tokens which is done sequentially and considered as a major bottleneck in degrading the performance of dataflow systems. The ADC eliminates the need for token generation and matching in dataflow systems and thus, by removing this bottleneck, the performance of the system improves.

In ADC, a dataflow graph is executed in two phases, namely the *search phase* and the *execution phase*. During the search phase, the dataflow graph can be assumed to be upside-down wherein the node(s) at the top-level is the *parent* and the nodes at the lower level, connected to the parent through arcs, are the *children*. The objective of the search phase is for each parent to search for its children. Since this search is conducted using content-addressable memory (also called associative memory, hence the name *associative dataflow*) which has parallel search capabilities, the search phase is conducted much faster compared to the sequential token-matching done in conventional dataflow paradigm. As a result of conducting the search phase, each node (i.e., the instruction) of the dataflow graph (i.e., the program) knows what its operands are (i.e.,

data), where they are located (i.e., location of data), and also the destination node for the result. During the execution phase, instructions are carried out as in traditional dataflow computer systems.

To better understand the concept of parent and children nodes, consider a simple dataflow graph to compute X = a+b+c+d (Fig.1). The search phase of the associative dataflow concept requires that the given dataflow graph be turned upside-down in order for each parent to search for its children. The inverted dataflow graph to allow progress of this search phase is shown in Fig. 2, wherein the node at the top (N3) is at level 0, and the nodes N1 and N2 are at level 1. Node at level 0, i.e., N3, is the parent of the nodes at level 1, i.e., N1 and N2, or in other words, the nodes N1 and N2 at level 1 are the children of the node N3 at level 0. Similarly, operands' pairs (a,b) and (c,d) are the children of the nodes N1 and N2 respectively.
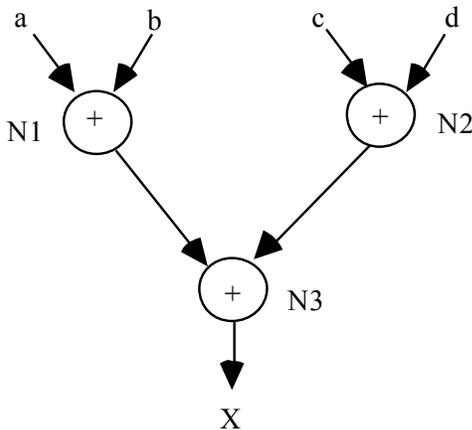
Fig. 1.   Dataflow graph to compute X=a+b+c+d
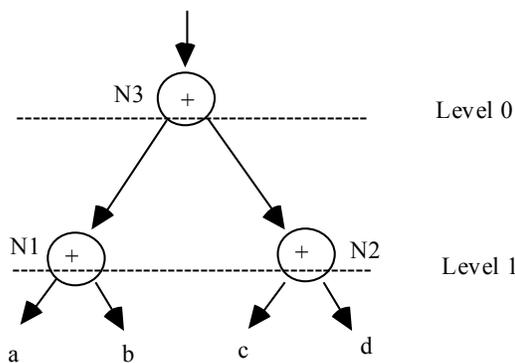
Fig. 2.   Dataflow graph to compute X=a+b+c+d inverted to allow progress of search phase

In traditional dataflow machine, execution of dataflow graph given in Fig. 1 would require 3 ALU Execution Times

(one for each node N1, N2, N3) plus 1 Token Matching Time (at node N3) plus 1 Memory Fetch Time (at Node N3). In associative dataflow machine, the Fig. 1 dataflow graph would need only Total Search Time + Total Execution Time. It has been shown in [11] that associative dataflow processor, based on ADC, has much better peak and benchmark performance figures compared to typical dataflow machines.

Considering that CBNS provides an efficient format for representing data and ADC exhibits a promising future for parallel processing, it was natural for researchers to consider amalgamating the two ideas into designing of a *Complex Binary Associative Dataflow Processor* (CBADP) which takes advantage of the best features found in both concepts.

IV.   COMPLEX BINARY ASSOCIATIVE DATAFLOW PROCESSOR (CBADP)

The schematic block diagram of a CBADP is given in Fig. 3 [12]. Each component of the diagram is described in the following sub-sections.

A. Associative Memory

An associative memory, also known as content-addressable memory (CAM), is defined as a collection of storage elements, called associative cells, which are accessed in parallel on the basis of data contents rather than by specific address or location. Each associative cell has the hardware capability to store and search its contents, in parallel, against the input data, and then indicate a match or mismatch by the state of a flip-flop. CBADP associative memory consists of a comparand register which contains the data to be compared against the contents of the memory array, a mask register used to mask off portions of the data word(s) which do not participate in the operations, a memory array containing a collection of memory cells providing storage and search medium for the data, and a responder indicating success or failure of a search operation.

There are two types of nodes in a dataflow graph, namely the *action-node* responsible for executing arithmetic or logic operation on the operands, and the *control-node* responsible for transferring the task of operation execution to some specific node when a certain condition is satisfied, as in case of branch operations. Depending upon the type of node, an associative memory word for action-node is of 80-bits length while a control-node word is of 18-bit length. Details about the format of the memory word can be found in [11]. Total number of words in the memory array are fixed at 64 to allow for up to 16 levels of dataflow graph with no more than four nodes per level.

B. Complex Binary Processing Unit

This unit is composed of four arithmetic and logic units to allow for up to four parallel operations corresponding to each node in a given level of the dataflow graph. Note that each CBALU is capable of handling arithmetic and logic operations in CBNS format, according to the algorithms described in Section II. Detailed designs of arithmetic circuits can be found in [13,14,15,16,17]. Currently, the instruction set of CBADP is composed of 21 instructions but there is a possibility of it to be extended to a total of up to 64 different instructions.
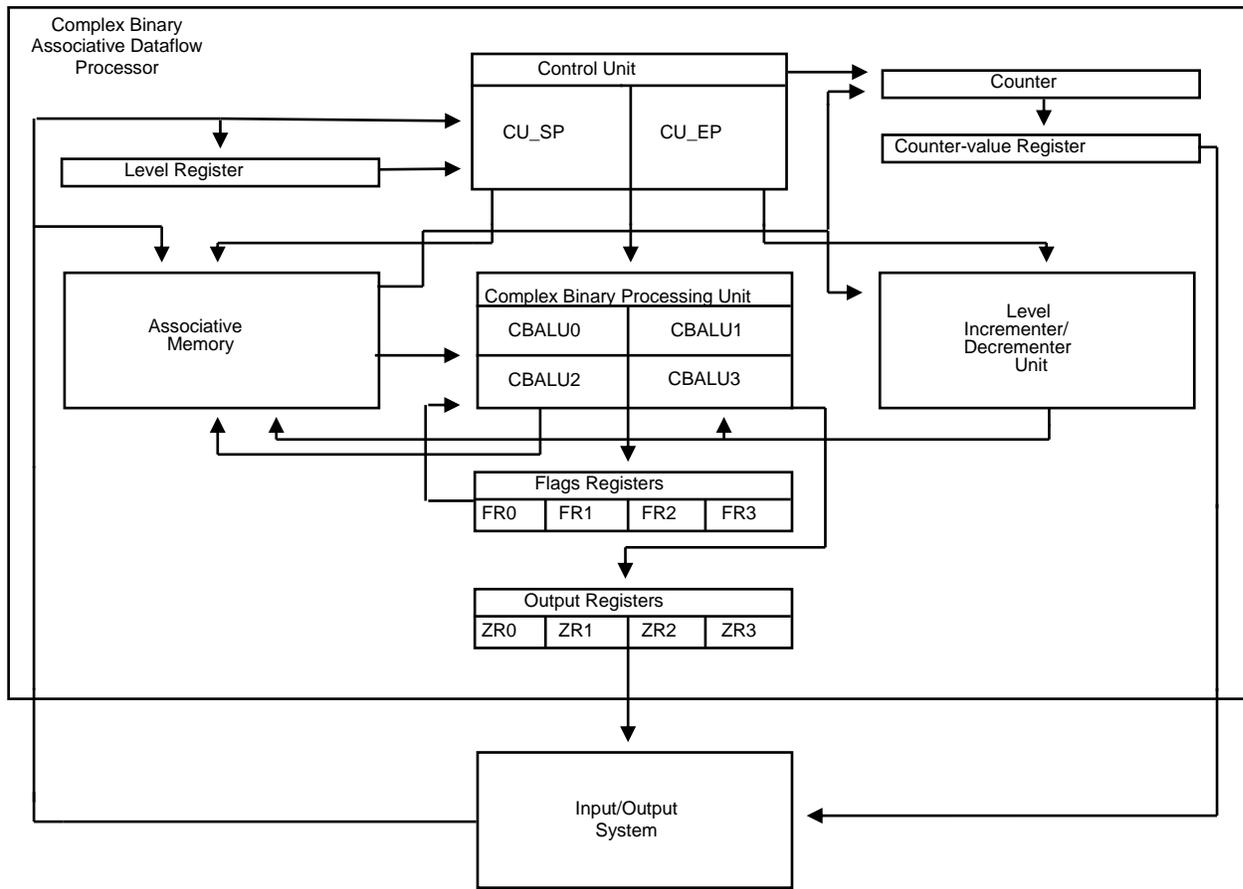
Fig. 3. Schematic block diagram of Complex Binary Associative Dataflow Processor (CBADP)

C. Level Incrementer/Decrementer Unit

To allow for 16 levels within a dataflow graph, a 4-bit level incrementer/decrementer unit is used to facilitate incrementing of the level number of the dataflow graph by one during the search phase and decrementing of the level number by one during the execution phase. Levels are labelled from top-to-bottom of the dataflow graph in descending order. Since, during the search phase, the graph is assumed to be upside-down, parent nodes are at the lower level and they are searching for children nodes which are at the higher level. This requires incrementing of the level number in the associative memory word as the search progresses. During the execution phase, the data is passed from nodes at higher level to the nodes at the lower level which requires decrementing of the level numbers as the execution phase continues. Given input E such that level incrementer/decrementer unit increments when E=0 and decrements when E=1, and current level number given by $X_0X_1X_2X_3$, the new level number $L_0L_1L_2L_3$ is given by the following Boolean equations:

. : Logic AND
+: Logic OR
X_i^: Logic NOT
@: Logic XOR

$$L_0 = E.[X_0^.X_1^.X_2^.X_3^ + X_0.X_2]$$
$$+ E^.[X_0^.X_1.X_2.X_3 + X_0.X_1^]$$

$$+ X_0.[X_1.X_2^ + X_1^.X_3 + X_2.X_3^] \quad (1)$$

$$L_1 = E.[X_1.X_2 + X_1^.X_2^.X_3^]$$
$$+ E^.[X_1.X_2^ + X_1^.X_2.X_3]$$
$$+ X_1.[X_2 @ X_3] \quad (2)$$

$$L_2 = E @ X_2 @ X_3 \quad (3)$$

$$L_3 = X_3^ \quad (4)$$

D. Control Unit

Independent hardwired control units for search phase (CU_SP) and execution phase (CU_EP) have been designed for CBADP. A Hardware Programming Language (AHPL) has been used to write a total of 24 control sequences for CU_SP and 23 control sequences for CU_EP. Details of control unit design are presented in [11].

E. Counter

This is a 6-bit up-down counter to keep track of up to 64 nodes which may be present in a dataflow graph to be executed on CBADP. Given input $x_1=0$ such that the counter counts up by one from its current state $x_2x_3x_4x_5x_6x_7$ to the next state $y_0y_1y_2y_3y_4y_5$, and counts down when $x_1=1$, the Boolean equations for the next state are given as follows:

. : Logic AND
+: Logic OR
$X_i^\wedge$: Logic NOT
@: Logic XOR

$$y_0 = x_1.\{x_2.[x_3 + x_5 + x_3^\wedge.x_4^\wedge.x_7^\wedge.(x_5^\wedge + x_6)]$$

$$+ x_2^\wedge.x_3^\wedge.x_4^\wedge.x_5^\wedge.x_6^\wedge.x_7^\wedge\}$$

$$+ x_1^\wedge.\{x_2.[x_3^\wedge + x_4^\wedge + x_3.(x_4.x_6.x_7^\wedge + x_5.x_6^\wedge)]$$

$$+ x_2^\wedge.x_3.x_4.x_5.x_6.x_7\}$$

$$+ x_2.(x_4 @ x_5) \tag{5}$$

$$y_1 = x_1.\{x_3.[x_4 + x_4^\wedge.(x_5^\wedge.x_7 + x_6.x_7^\wedge)]$$

$$+ x_3^\wedge.x_4^\wedge.x_5^\wedge.x_6^\wedge.x_7^\wedge\}$$

$$+ x_1^\wedge.\{x_3.[x_4^\wedge + x_5.x_6^\wedge + x_4.x_6.x_7^\wedge]$$

$$+ x_2.x_4.x_5.x_6.x_7\}$$

$$+ x_3.(x_4 @ x_5) \tag{6}$$

$$y_2 = x_1.(x_4.x_6 + x_4^\wedge.x_5^\wedge.x_6^\wedge.x_7^\wedge)$$

$$+ x_1^\wedge.(x_4.x_5^\wedge + x_4^\wedge.x_5.x_6.x_7)$$

$$+ x_4.(x_5.x_6^\wedge + x_5^\wedge.x_7 + x_6.x_7^\wedge) \tag{7}$$

$$y_3 = x_1.(x_5 @ x_6^\wedge.x_7^\wedge)$$

$$+ x_1^\wedge.\{x_5.x_6^\wedge + x_6.[x_5 @ x_7]\} \tag{8}$$

$$y_4 = x_1 @ x_6 @ x_7 \tag{9}$$

$$y_5 = x_7^\wedge \tag{10}$$

F. Counter-value Register

This register (6-bit) is used to store the counter value at the completion of each successful search phase. This information is then used by the operating system in formulating an efficient execution phase for the given dataflow graph.

G. Level Register

This 4-bit register holds information about the maximum level number in the given dataflow graph. The control unit uses this information to ensure that all levels of the graph have been searched and executed.

H. Flags Registers

There are four of these registers corresponding to each CBALU in the complex binary processing unit. Each register holds carry, zero, negative, and overflow flags.

I. Output Registers

There are four of these registers corresponding to the output

produced by each CBALU. These registers are accessed by the Input/Output System which forms the interface to the outside world.

V. CBADP SOFTWARE DESIGN ISSUES

There is a natural communication gap between man and machine. Computer hardware operates at a very atomic level in terms of bits and bytes, whereas people tend to express themselves in terms of natural languages such as English or in mathematical notation. This communication gap is bridged by means of an artificial language which allows the man to express himself with a well-defined set of words, sentences, and formulas that can be "understood" by a computer.

To achieve this interaction, the human is supplied with a user's manual which explains the constructs and meanings allowed by the language, and the computer is supplied with the software by which it can take a stream of bits representing the commands or programs written in the language by the human and translate this input into the internal bit patterns required to carry out the human's intent.

A. Compiler

A compiler takes as input a source program and produces as output an equivalent sequence of machine instructions. This process is so complex that it is not reasonable, either from a logical point of view or from an implementation point of view, to consider the compilation process as occurring in one single step. Not long ago, compilers were considered almost impossible programs to write. The design of a compiler for CBADP is required to follow the same phases as the ones used today for the design of any other type of compiler, namely lexical analysis, syntax analysis, intermediate code generation, code optimization, and then code generation. The programmer will be able to write programs in one of the high-level languages, such as C, Pascal, or Fortran, and then using the CBADP compiler, the source code will be converted to the machine instructions (in terms of 0 and 1), which will be fed to the hardware components within the complex binary associative dataflow processor for onward processing and execution.

B. Operating System

An operating system is a program that acts as an intermediary between the user of a computer and the computer hardware, and its purpose is to provide an environment in which the user can execute programs [18]. An abstract view of a CBADP system is given in Fig. 4.

The CBADP provides the basic computing resources (associative memory, processing unit, registers) which are managed by the CBADP operating system, according to the user programs, which are written in some high-level language (such as C, Fortran, or Pascal) and have been converted into machine language using the CBADP compiler.

Fig. 4.   Abstract view of CBADP system

The CBADP operating system has several important functions for smooth and efficient processing of the user programs, and these are:

- Based on the information provided by the CBADP compiler about the given dataflow graph, design an efficient search phase for the program.

- Provide the system call to the CU_SP (control unit for search phase) for initiating the search phase of the dataflow graph.

- Acknowledge and handle any interrupts generated during the search phase to keep track of the information about the nodes (of dataflow graph) and their specific locations within the associative memory where they have been stored.

- Based on the information gathered during the search phase, design an efficient execution phase for the given dataflow graph.

- Provide the system call to the CU_EP (control unit for execution phase) for initiating the execution phase of the dataflow graph.

- Acknowledge and handle any interrupts generated during the execution phase to keep track of the information about the nodes (of dataflow graph) and their execution status in the system. In case of branch operations, enable or inhibit the appropriate control-node depending upon whether the branch condition has been satisfied or not.

- Provide synchronization mechanism between the four arithmetic and logic units contained within the processing unit of the CBADP so that the instructions are executed in proper order.

For best CBADP performance, the ideal operating system should have multitasking capabilities so that more than one dataflow graph can be processed at a given time if the necessary resources, such as CBALUs and registers, are available. For example, to achieve maximum utilization of the complex binary processing unit, all four of the CBALUs must be kept busy all the time. If there are less than four nodes at any given level for execution, the multitasking operating system can assign the idle ALU(s) to another dataflow program which can run in parallel with the current dataflow program. In case of more than one dataflow program being processed on the ADP, the protection becomes an important issue to be addressed. That is, steps need to be taken to ensure that the proper execution of one graph does not interfere with the proper execution of the other dataflow graph(s) in the system.

VI.   CONCLUSIONS AND FURTHER RESEARCH

The design of a content-addressable memory-based associative dataflow processor, for which an Australian Innovation Patent has been granted, has been presented in this paper. Work on establishing complex binary number system as a natural enhancement of traditional binary number system is still in progress and the given complex binary associative dataflow processor design is still undergoing improvements and implementation optimization for field-programmable gate arrays (FPGAs). An avenue of further research in this area would be to incorporate CBNS in an image processor design or a digital signal processor design, two of the well-known areas where complex numbers are widely used, and then compare their performances with the processors implementing base 2 binary number system. Incorporating cache memories within the CBADP and then determining statistical performance analysis of such a system will be a valuable contribution to this area of engineering. These days, efforts are underway to design computer systems which mimic human brain. Since human brain behaves very much like associative memory, a content-addressable parallel processor system, such as CBADP, may provide a useful tool for research in this area of interest to both medicine and engineering. The design of an operating system for CBADP is another interesting area of further research in this realm of computer engineering.

REFERENCES

[1]   D. Knuth, "An imaginary number system," Communications of the ACM, 1960, pp. 245-247.

[2] W. Penney, "A numerical system with a negative base," Mathematical Student Journal, May 1964, pp. 1-2.

[3] W. Penney, "A binary system for complex numbers," Journal of the ACM, April 1965, pp. 247-248.

[4] V. Stepanenko, "Computer arithmetic of complex numbers," Cybernetics and System Analysis, 1996, Vol. 32, No. 4, pp. 585-591.

[5] T. Jamil, N. Holmes, and D. Blest, "Towards implementation of a binary number system for complex numbers," Proceedings of the IEEE SoutheastCon 2000, Nashville, Tennessee (USA), April 7-9, 2000, pp. 268-274.

[6] T. Jamil, "The complex binary number system – basic arithmetic made simple," IEEE Potentials, December 2001/January 2002, Vol. 20, No. 5, pp. 39-41.

[7] T. Jamil, Complex Binary Number System – Algorithms and Circuits, Springer, 2013.

[8] D. Blest and T. Jamil, "Efficient division in the binary representation of complex numbers," Proceedings of the IEEE SoutheastCon 2001, Clemson, South Carolina (USA), March 30-April 1, 2001, pp. 188-195.

[9] D. Blest and T. Jamil, "Division in a binary representation for complex numbers," International Journal of Mathematical Education in Science and Technology, 2003, Vol. 34, No. 4, pp. 561-574.

[10] http://pericles.ipaustralia.gov.au/ols/auspat/applicationDetails.do?applicationNo=2010100706

[11] T. Jamil, Introduction to Associative Dataflow Processing – From Concept to Implementation, VDM Verlag, 2010.

[12] T. Jamil, "Design of a complex binary associative dataflow processor," Proceedings of the 4th International Conference on Computer Engineering and Technology, Phuket, Thailand, May 27-28, 2012, pp. 32-35.

[13] T. Jamil, B. Arafeh, and A. Al-Habsi, "Hardware implementation and performance evaluation of complex binary adder designs," Proceedings of the 7th World Multiconference on Systemics, Cybernetics, and Informatics (SCI 2003), Orlando, Florida (USA), July 27-30, 2003, Vol. II, pp. 68-73.

[14] J. Goode, T. Jamil, and D. Callahan, "A simple circuit for adding complex numbers," WSEAS Transactions on Information Science and Applications, July 2004, Vol. 1, No. 1, pp. 61-66.

[15] T. Jamil, A. Abdulghani, A. Al-Maashari, "Design of a nibble-size subtractor for (-1+j)-base complex binary numbers," WSEAS Transactions on Circuits and Systems, July 2004, Vol. 3, No. 5, pp. 1067-1072.

[16] T. Jamil, A. Al-Maashari, and A. Abdulghani, "Design and implementation of a nibble-size multiplier for (-1+j)-base complex binary numbers," WSEAS Transactions on Circuits and Systems, November 2005, Vol. 4, No. 11, pp. 1539-1544.

[17] T. Jamil and S. Al-Abri, "Design of a divider circuit for complex binary numbers," Proceedings of the World Congress on Engineering and Computer Science 2010/International Conference on Circuits and Systems, San Franscico, California (USA), October 20-22, 2010, Vol. II, pp. 832-837.

[18] A. Silberschatz and P.B. Galvin, Operating System Concepts, Addison-Wesley Publishing Company, 1994.

User Satisfaction Determinants for Digital Culture Heritage Online Collections

Zaihasriah Zahidi

Faculty of Creative Multimedia,
Multimedia University,
Cyberjaya, Selangor, Malaysia

Yan Peng Lim

Faculty of Creative Multimedia,
Multimedia University,
Cyberjaya, Selangor, Malaysia

Peter Charles Woods

Faculty of Creative Multimedia,
Multimedia University,
Cyberjaya, Selangor, Malaysia

Abstract—The aim of this paper is to identify the possible determinants that influence user satisfaction in the context of digital cultural heritage (DCH) online collections. The data was collected in 3 stages. For the first stage, literature studies were conducted in understanding the general overview about user satisfaction in various web-domains. Next, think-aloud protocol was conducted with a group of general user with nil background of cultural heritage. Two existing digital culture heritage online collections were used as the vehicle to get the findings. Lastly, existing studies on Herzberg Two-Factor Theory in web-environment context was adapted and adopted in identifying the possible hygiene and motivator factors which influence the user satisfaction in this context of study.

Keywords—User experience; user satisfaction; digital culture heritage online collections

I. INTRODUCTION

User satisfaction is one of the components in human computer interaction. According to Merriam Webster [1], satisfaction is defined as "the act of providing what is needed or desired". In any interaction between a user and an interface, achieving user satisfaction is the key in determining the successful of a product or a system Alawneh, Al-Refail and Batiha [2], user satisfaction is subjective to measure. Factors that influence a user to feel satisfy is unique to one another depending on individuals' needs, expectations and existing experience when interacting with an interface [3]. Generally in web environment, satisfied user may "spend longer at a website, may revisit the website later, and may recommend the website to others [4]. Thus, investigating possible website features of a web interface contribute to the satisfaction would be useful to study.

Despite vast area of studies had been conducted about user satisfaction in web environment [5][6][7], the studies related to user satisfaction in digital cultural heritage (DCH) online collections domain are limited [8]. UNESCO defined DCH as "…made up of computer-based materials of enduring value that should be kept for future generations". DCH online collections are usually online-based repositories of digitised cultural heritage assets [9] as a means of preservation. Preservation is nothing without public access. Hence, many major museums nowadays are going digital by digitising physical exhibit items and publish them online for worldwide audience.

The lack of findings on what makes user satisfies and dissatisfies specifically in DCH online collections may result to poor user experience (UX). User satisfaction is the result of good UX. The goal of UX is to create an overall positive experience for the user through the utility, ease of use and pleasure provided when interacting with an interface [10]. Europeana in its Strategic Plan 2011 – 2015, has recognized the importance of user satisfaction in DCH. In Malaysia, one of the agendas in National Policy of Creative Industry [11] is to urge user satisfactory studies towards the digital content of cultural heritage.

The aim of this paper is to identify the possible determinants that influence user satisfaction in the context of DCH online collections. The objective of this paper is to identify possible DCH web features that could be user satisfaction determinants. The determinants are important to be identified in understanding what makes such online collections produced are meeting the users' needs, expectations and existing experience. It is to provide insights that DCH online collection is not just a means of cultural heritage preservation but also is significant to the human-computer interaction context. This study uses existing studies of Herzberg's Two-Factor findings as guidance in assisting the identification of DCH web features.

Herzberg's Two Factor Theory is a motivation theory based on two factors, which are motivator and hygiene factor. The theory derived based on studies about factors that lead to workers' satisfaction (motivator) or dissatisfaction (hygiene) in a working environment. Motivator factors are fulfilled by intrinsic feeling such as work achievement, job recognition, work itself, level of responsibility, advancement and growth. On the other hand, hygiene factors are the jobs' basic needs such as company policies, supervision, working conditions and salary [12]. If these needs are not fulfilled, workers tend to feel dissatisfied.

In referring to Herzberg's theory, Zhang, Small, von Dran and Barcellos [4] proposed that creating a motivating website is similar to create a motivating workplace. In web environment, hygiene factors consists of the functionality of the website feature. For instance, a search feature in a website is useful in assisting user to seek for information by keywords. With such feature, it complements user-searching behaviour beyond clicking on navigations links or buttons. If such feature is absent, it might cause user dissatisfaction in seeking information by searching using keywords. Motivator factors in a website can be seen as factors that enhance user satisfaction. It is beyond the hygiene factors which might be subjective to the users. For instance, the usage of multimedia elements in a

website to attract user. The usage might impress user and hence increase the overall user satisfaction. However, if such feature is absent, user might not feel dissatisfied and only will leave user a neutral feeling towards it. Both hygiene and motivator go hands in hands. The hygiene factors must be present or else users will feel dissatisfied but if motivator factors are absent, it might leave user with neutral feeling but dissatisfied as long as the hygiene factors are fulfilled.

Questions that directed this study are:

1) What are the user satisfaction determinants of DCH online collections?

2) How would Herzberg Two-Factor Theory be applied to determine the user satisfaction determinants?

II.   USER SATISFACTION IN VARIOUS DOMAINS

General review of user satisfaction in various domains was conducted in this paper. The definitions of user satisfaction and factors to influence it according to different domains were gathered to make it more comprehensible. The five (5) different domains that are selected are end-user computing, websites, e-satisfaction, e-retailers/e-services, online digital libraries and online tourism websites. Although all domains are web-based, each has different purposes and functions. Because of the differences, it is essential to study and identify what are the common definitions and factors to influence user satisfaction. This is to get the general overview of user satisfaction for web-based environment. Table 1 provides the definitions of user satisfaction in various web domains.

With the many definitions of user satisfaction being defined differently in different contexts, it can be seen that user satisfaction is subjective to measure and unique depending on the purpose of the interface or system. It is depending on the users' type, purpose and needs in interacting or using the interface or system.

TABLE.I.   USER SATISFACTION DEFINITIONS IN VARIOUS WEB DOMAINS

Domains	Definitions
End-user computing	"…is defined as the opinion of the user about a specific computer application which they use" [13]. "the extent to which users believe the information system available to them meets their information requirements" [14]. "a perceptual or subjective measure of system success" [15].
Websites	"stickiness and the sum of all the website qualities that induce visitors to remain at the website rather than move to another site" [16]. It relates to the user's attitude about the website – how enjoyable it is to use it [17].
E-Satisfaction	Ability for a service portal to be compatible with citizens' needs, desires and expectations [2].
E-Retailers/E-Services	"Customer satisfaction means how a company provides, supplies or deliver products or services to meet customer needs and wants" [18].

III.   FACTORS TO INFLUENCE USER SATISFACTION

There are many factors influencing user satisfaction. Usability, learnability, functionality, accessibility and ease of use are the factors to influence overall user satisfaction [19]. Aesthetic, [20] interface design and joyful of use also could influence the factors [21]. Table II below displays the summaries of the factors to influence user satisfaction according to 6 different domains.

TABLE.II.   FACTORS TO INFLUENCE USER SATISFACTION ACCORDING TO VARIOUS DOMAINS

Domain	Authors	Factors to Influence
1. End-user computing	[22]	"Content, accuracy, format, ease of use, timeliness, satisfaction with system's speed, system reliability in End-User Computing Satisfaction (EUCS) that influence most end-users' satisfactions"
2. Website Design	[23]	Site organization, information content and navigation and revisit the website.
	[24]	Visual design, information architecture, information design, navigation design, content and interaction design
3. E-Government Services	[2]	• Security and privacy • Trust • Accessibility • Awareness of public services • Quality of public services
4. Online Tourism Websites	[25]	• Functionality • Usability
	[26]	• Information and process • Value added • Relationships • Trust • Design and usability
	[27]	• Ease of use • Joy of use

		• Content • Interactivity • Transaction support • Added value • Appearance • Clear navigation paths
	[28]	• Information quality • Security • Website functionality • Customer relationships • Responsiveness
	[29]	• Interface • Perceived quality • Value
	[30]	• Click Stream Paradox • Security Value Information • Accuracy • Interactivity • Loading Speed • Purchase Influence Recommend-ability
5. Online Shopping Websites	[31]	• Convenience • Merchandising • Security • Serviceability
	[32]	• Navigation • Usefulness • Convenience • Ease of use • Sub-experience (substitutability) • Enjoyment (interactive elements with users)
	[33]	**Computer factors** • Neat interface • Consistent web design • Updated information • Security in payment method **Human factors** • Global search feature • Humor • Links to similar websites • Feedback features • Visitors count **Entertainment** • Enjoyable • Pleasing • Entertaining **Informativeness** • Provide resourceful and relevant information **Irritation** • The website is frustrating **Usefulness** • The website can improve shopping performance, productivity and effectiveness **Attitude** • Feel satisfy with the service provided **Flow** • Judging the website as interesting, fun, exciting and enjoyable **Purchase Intentions** • Intent to purchase soon

		Revisit Intentions • Intent to revisit soon
	[34]	• Usability • Information quality • Visual appeal
6. Digital Online Libraries	[35]	• User friendly interface • Simple interface to access rich information maintained by cataloguers for decades
	[36]	Ease of use predominantly by having: • Use clear and simple terminologies and instructions to be understood by general users • Consistent interface Replacing text buttons with graphical icons to increase user's attention • Easy navigation to reduce cognitive effort for information searching
	[37]	Users prefer: • Updated content • Structured and leveled information presentation • Easiness of user to discover information • Learnability
	[38]	Social metadata features; Information contributed by users through: • Tagging • Comments • Reviews • Ratings • Recommendations evaluate the content.

IV. METHODOLOGY

The study of this paper was conducted in 3 stages. Below are the details of each phase:

Stage 1

The objective of Stage 1 was to have a general overview about user satisfaction specifically in web-based environment domain. The definitions and factors to influence user satisfaction were reviewed based on literature and existing studies. Various domains were selected including end-user computing, website design, e-government services, online tourism websites, online shopping websites and digital online libraries. Each domain differs from one another in terms of purpose and functions. Despite the differences, it is important to discover the common factors in influencing user satisfaction for web-based platforms in general.

18 papers inclusive of different domains were chosen to study the factors. Factors of each paper were accordingly listed. Word frequency and thematic analysis were used to categorise similar categories that represent common meaning. Based on the word frequency, similar words that represent similar meanings were coded and frequencies of words were recorded. Twelve (12) themes were emerged during the analysis stage by using thematic analysis. The themes were listed from the most rated frequency to the least according to the chosen literature studies.

The themes were:
1. Content and information
2. Interface & consistency
3. Website functions and features that promote ease of use
4. Easy navigation
5. Security
6. Positive feelings towards the website
7. Value added
8. Accuracy
9. Revisit website
10. Trust to the resources
11. Relationship among users
12. Accessibility

Stage 2

Data collected in this stage was to explore whether existing DCH online collections are able to satisfy general type users. Two (2) existing DCH online collections were used as the vehicle to get the findings as shown in Figure 1 and Figure 2. Both were labeled as Website A and Website B. Website A contains Malaysian culture heritage whilst Website B is more global but mainly about the Western content.

Website A

A non-commercial and approved concept of online metadata archive platform for Malaysian culture and heritage content. Users are able to contribute content that is related in forms of photographs, videos, links and oral stories to the website upon the website admistration approval. No user account is required in order to use this website.

Fig.1. Malaysian Culture & Heritage Digital Bank [39]

Website B

A commercial user-generated of global content which display historical images with the concept of "pinning

photographs on the map". User needs to create an account in order to contribute photos but are free to browse through the collections as guest.

Fig.2.   Historypin [40].

The criterion of the chosen websites were 1) public-accessed online platform 2) purpose of website is to archive cultural heritage content 3) interface and approach in displaying content of both website must be dissimilar between the two websites. The relevance of these would help the authors to explore more in understanding the factors that could influence user satisfaction in the context of this study.

A purposive sampling was adopted in the study. Fraenkell and Wallen [41] defined purposive sampling as "a non random sample selected because prior knowledge suggests it is representative, or because those selected have the required information". 14 participants were involved in this test, consisting of 8 undergraduate students, 3 postgraduate students and 3 academic researchers. These participants represented the general users who were not engaged with cultural heritage background.

Think-aloud protocol was applied with purposive sampling. Samples were asked to browse two (2) DCH online collections websites and were assigned to complete the tasks given. The gist of the tasks was searching for information by using the websites' user interface. Tasks created were based on the UX component suggested by Hartson and Pyla [42] which are usability, usefulness, emotional impact during interaction and savouring the memory after interaction. Oral data were then recorded and transcribed. Appropriate data and evidence recorded in Stage 2 with the identified themes in Stage 1 were taken into consideration in suggesting possible hygiene and motivator factors.

Results indicated that general users showed satisfaction when a website provides:

1) Attractive layout and visuals with dynamic interactivity that captures users attention.

2)   Content are organised neatly and brief descriptions to assist users with information seeking behaviour.

3)   Usability of the interface is expected to be there but it will not influence their overall user satisfaction.

Stage 3

The objective of this stage was to identify possible hygiene and motivator factors of DCH online collections. The identification were referred to the findings contributed by Zhang, Small, von Dran and Barcellos [4] as it suggested website features that could be able to provide user satisfaction based on Herzberg Two-Factor Theory. In Stage 3, the themes emerged in Stage 1 and think-aloud protocol evidences in Stage 2 were used to identify possible hygiene and motivator factors in DCH online collections setting.

The stages involved in this study are summarised as in Figure 3 below.
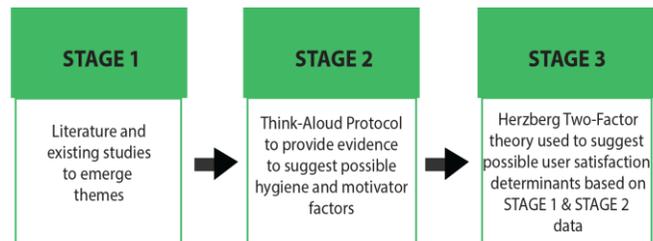
Fig.3.   Stages involved in this study.

V.   FINDINGS

A. Possible User Satisfaction Determinants of Digital Cultural Heritage Online Collections

The results of this study referred to Zhang, Small, von Dran and Barcellos [4] findings in determining websites features that could be able to satisfy users according to Herzberg Hygiene and Motivator Factors. The theorized examples of possible both hygiene and motivator features in DCH online collections environment were suggested based on the think-aloud-protocol findings and secondary data. Table III shows the possible hygiene and motivator features in DCH online collections:

TABLE.III.   POSSIBLE HYGIENE FEATURES IN DIGITAL CULTURAL HERITAGE ONLINE COLLECTIONS

Herzberg's Hygiene Factors	Specific Example of Herzberg's Hygiene Factors	Theorized Application to the Web Environment	Theorized Examples of Possible Hygiene Features in Web Environment	Theorized Examples of Possible Hygiene features in Digital Cultural Heritage Online Collections Environment
Working Condition	Light, temperature, furniture, office size, "tools or equipment" to get tasks done, first impression or general appearance	First impression or general appearance	1. Brightness of the screens/pages 2. Utilization of the screen size (viewable size of the screen) 3. Screen background color and pattern 4. Sharpness of displays (including images) 5. Eye catching image(s) or title on the homepage that makes you want to continue exploring the site	1. Overall layout is simple & neat 2. Overall layout is consistent in every pages 3. Color & design theme is consistent 4. Color & design theme is bold and attractive 5. Interface designed creates the expectation of "fast information searching" 6. Website is responsive regardless different screen devices
		Basic functions/features that help to get tasks done	6. Live/broken links 7. Consistent use of link colors within the web site 8. Existence of unloadable items that are not central to the task (e.g. non-found images are used as bullets or decoration) 9. Need to scroll to view the homepage 10. Need to scroll to view the detailed/content pages 11. Robustness of the web interface (user mistake-tolerant, few bugs) 12. Stability of the site: should be consistently available for access 13. Support for different platforms and/or browsers 14. Search function/engine to work with large amount of info on the web site	1. Search feature 2. Tag feature 3. Comment feature 4. Geo-location feature 5. Google Street View feature 6. Account Login
Company policy and administration	Procedures or rules of doing things; pace of feedback from administration; privacy and proper use of employee's private information; in general the bureaucratic aspects of the working environment	Requirements for doing tasks	15. Length of the procedure to complete a task (e.g. steps/pages/actions to go through in order to get certain info) 16. Time on learning to use and becoming skillful at using the site	1. Fewer clicks to get information needed
		Feedback or response	17. Length of a page's loading or responding time 18. Indication of system action time expectation (e.g. long loading time warning)	1.Faster content loading 2.Notify user current status of action (e.g: loading time)
		Access restriction	19. Access restrictions (e.g. one needs to pay a fee, to sign on, to enter a password, or to provide some private info before one can access task-related info)	1. Account login is needed to secure activities performed 2. Automatic account login with social media account
		Privacy and data confidentiality	20. Collection of user's data without user's knowledge (including using cookies, write to user's local machine) 21. Informing users that their information will be collected 22. Declaration of specific use of the information that users need to provide (e.g. declare for statistics only, not to provide to the venders, not for marketing purpose, etc.)	1.Deposited data by users need to be original

Interpersonal relations	Co-workers attitudes, perceptions and trust	Credibility of owners/designers and the website: trust and trustworthy	23. Identification of site owners/designers 24. Credibility of the website owner/designer 25. Credibility of the website (e.g. the site won awards) 26. Number of times the website has been visited (e.g. shown by a counter)	1.Content is certified & verified by authority 2. Credibility of the websites (e.g: the site won awards) 3. Visitors counts 4. Familiar content will attract users to browse
Interpersonal relations	Co-workers attitudes, perceptions and trust	Web owners/designers' attitudes and perceptions	27. Information about improper or controversial materials 28. Indications of gender or racial/ethnic biases and stereotypes	1. Disclaimer about the royalty of the content provided as a user-generated content
Supervision	Authority; guidance & support; availability of the supervisor; technical support	Authority and availability of owners/designers	29. Authority of the web designer/owner 30. Indication of the purpose or objective of the web site or potential audience 31. Availability of designer/owner for further information (e.g. email)	1.Clear purpose of website is presented to the user 2. FAQ provided for users 3. Direct communication between user and website by providing email, address, person in charge 4. Expand communication by suggesting to Add in social media accounts (e.g: Facebook, Twitter, Google Plus, LinkedIn)
		Navigation	32. Working navigation aids (buttons or links) where necessary 33. Be able to know where to get started with the site's primary features 34. Be able to determine current position within the site 35. Simple and clear directions for using the website	1. Navigation is simple 2. Navigation is straightforward 3. Navigation that reduce confusion 4. Navigation with less clicks 5. Navigation that obvious 6. Navigation enhances effective information searching

TABLE.IV.   POSSIBLE MOTIVATOR FEATURES IN DIGITAL CULTURAL HERITAGE ONLINE COLLECTIONS.

Herzberg's Motivation Factors	Specific Example of Herzberg's Motivation Factors	Theorized Application to the Web Environment	Theorized Examples of Possible Motivation Features in Web Environment	Theorized Examples of Possible Motivation features in Digital Cultural Heritage Online Collections Environment
Work Itself	Work-related tasks are challenging, stimulating, interesting, meaningful, useful, creative, fun	The information seeking tasks	36. Interestingness of the browsing task 37. Challenge of the browsing task 38. Usefulness of the browsing task to job/work, school, etc. 39. Meaningfulness of the browsing task 40. Fun to explore	1. Interaction is highly interactive 2. Content delivered in non-formal way 3. More visual in delivering about content 4. Use interesting, never seen shots 5. Provide updated content 6. Avoid delivering common content
	Quality of the information content: what a website covers ** (relevant, timely and current, complete and accurate, objective and novelty, understandable, consistent)		41. Task-relevant information 42. Relevant links (to the task, context, or information content) 43. Amount of irrelevant information (such as online ads, meaningless images) 44. Up-to-date information 45. Indication of addition of new information in the future 46. Complete/comprehensive/inclusive/adequate coverage of information 47. Precise/accurate and referenced information 48. Objective, unbiased information 49. Indication of limitations of information (e.g. source, coverage, date last modified) 50. Novelty and interesting information 51. Understandable information 52. Appropriate detail level 53. Coherent content that supports the web site's intended purpose/objective	1. Brief information to deliver content 2. Detailed information is presented by choice (e.g: click here to know more) 3. Extra information is presented via external links 4. Present user with the numbers of items in the collections 5. Present user with the latest addition in the collections 6. Present user with the latest contributor of the item 7. Brief description to describe an item should be standardised 8. Standard metadata is used to described items 9. Information delivered should be understandable for all range of age

		Presentation/organization on of information: how a website covers information ** (information architecture, aesthetic and affective, learning consideration)	54. Logical organization of information within the website (e.g. by topic, by date, from broad to narrow) 55. Familiar terminology 56. Consistent use of terms and graphics 57. Overview, table of contents, summaries/headings 58. Scannability of a page (incl. chunks, screen uncluttered, highlights, etc.) so that users can easily scan the page to get info without reading line by line 59. Visually pleasing screen layout 60. Visually pleasing color use 61. Multimedia that adds information value 62. Variety of media (audio, video), formats (visual oriented or analytical oriented), types (use of examples, questions, plain descriptions) for different learning or thinking styles 63. Use of humor	1. Content is presented in media-rich form of information 2. Content is presented in visually engaging way 3. Content is presented in interactive way 4. Content is presented précised & briefly 5. Detailed content is presented optionally 6. Content is presented with high quality of photographs 7. Content is presented with rare and interesting collections of photographs 8. Content is presented with additional information using external links 9. Content is presented in comprehensive manner 10. Content is presented clearly 11. Content is structured effectively based on information hierarchy 12. Content is categorised using proper taxonomy 13. Content is indexed using standard indexing system 14. Content is tagged properly
Achievement	Successful task completion.	Task completion	64. Achieved results for the task 65. Quality of the task results 66. Time spent on the task 67. Problems solved (e.g. users may encounter unexpected problems while conducting the task and eventually solve the problems)	1. Searchable is important because once a keyword is not searchable, frustration appears 2. Search suggestion to assist search activity 3. Broken links should not make exist to avoid frustration
Responsibility	Certain control or power over the environment; make job related decisions with a minimum supervision	User control	68. User control of amount of information accessed 69. User control of procedures/steps of accessing information 70. User control of difficult levels (or details/depth) of information to be accessed (e.g. headings and details in a page allows a user to decide to either read heading or go for more details) 71. User control of pacing (how fast to go through the website) 72. Opportunities for interactivity	1. User has the authority to contribute content 2. User has the authority to comment on content 3. User has the authority to remove contributed content
Advancement & Growth	Professional advancement; Growth potential in task capability, knowledge or skills	Knowledge or skills gained	73. New skills, knowledge gained by doing the tasks on the website	1. Feel empowered when discovering new information 2. Feel proud after discovering familiar local stories
Recognition	Recognition by peers or supervisors for performance; real skills and capacities are put to use on jobs	Recognition by owners/designers on knowledge and skill levels	74. Assumed/recognized audience's knowledge and skill levels	1. Reward active contributor with online recognition

VI. Discussion

The study suggests that the user satisfaction determinants for DCH were identified supported with literature studies and according to user's evidences. These findings would be useful as a stronger guidance for designers to present the content of a cultural heritage online collections in a website that satisfies user by considering the user satisfaction determinants.

Interesting data from Stage 2 suggested that general users which are students associate user satisfaction with the overall vibrant and bold interface of the website while general users which are the academic researchers link user satisfaction with the quality of content provided in terms of information architecture and comprehensiveness of the DCH content. In terms of interactivity and advanced use of interaction in such websites, general users (students) highly appreciated it and gives them a sense of joyfulness to explore the website more with the interactivity. In contradiction with the general user (academic researchers), majority ignored any advanced

interaction. Such features did not influence them to be satisfied with the website due to occupied daily tasks and prefer the standard way of web-based interaction with fewer clicks. Although users are the same user type, they possess different needs and demands. For instance, students have the willingness to explore the website more with the advanced type of interaction provided while the academic researchers refuse to explore further due to busy work demand. These are the important aspects to be understood in examining what makes user satisfies or dissatisfies when interacting with DCH online collections.

Existing studies on Herzberg's Theory were adapted and adopted in determining the user satisfaction determinants. Zhang, Small, von Dran and Barcellos [4] findings indicated examples of both hygiene and motivators of web features in a general context of web-based environment. Based on these fundamental and understanding, this study had extended existing studies and interpreted it in the context of DCH. With the data from literature studies and think-aloud-protocol, the validity of the findings can be assured.

VII. RECOMMENDATION AND FUTURE WORK

Results of this study were based on a small-scale sample and only consider a single type of user which is the general user with nil background of cultural heritage. For future work, it is recommended to apply this study to users with cultural heritage background. Different sample of background would give different perspective. This can be useful as an additional guidance to the designers to consider when designing a DCH online collections websites.

VIII. CONCLUSION

In conclusion, possible determinants of user satisfaction in the context of DCH online collections website were identified in this study. Although the findings represent the general users, however it still considers users different needs for both students and academic researchers. It is difficult to design a website that meets all types of users' needs but at least it can be done by fulfilling the general determinations of user satisfaction.

REFERENCES

[1] Merriam-Webster.com. 2014, http://www.merriam-webster.com/dictionary/satisfaction

[2] A. Ali, H. Al-Refai, K, Batiha, "Measuring user satisfaction from e-government services: Lessons from Jordan," Government Information Quarterly, 2013, 30(3), pp. 277-288. [Online]. Available: 10.1016 / j.giq.2013.03.001. [Accessed: Aug. 10, 2014].

[3] Zaihasriah Z., L. Yan Peng, P. C. Woods, "Understanding the User Experience (UX) Factors that Influence User Satisfaction in Digital Culture Heritage Online Collections for Non-Expert Users," Science and Information (SAI) Conference, London, August, 27-29, 2014.

[4] P. Zhang, R. V. Small, G. M. von Dran, S. Barcellos, "Websites that Satisfy Users: A Theoretical Framework for Web User Interface Design and Evaluation," Proceedings of the 32nd Hawaii International International Conference on System Sciences, Maui, Hawaii, January, 5-8, 1999.

[5] P. Zhang, R. V. Small, G. M. von Dran, S. Barcellos, " A Two Factor Theory for Website Design," Proceedings of the 33rd Hawaii International International Conference on System Sciences, Maui, Hawaii, January, 4-7, 2000.

[6] L. Xiao, S. Dasgupta, "Measurement Of User Satisfaction With Web-Based Information Systems: An Empirical Study," Proceedings of the 2002 Americas Conference on Information Systems, August 9-11, 2002, Dallas, TX.

[7] T. W. Zazelenchuk, E. Boling, "Considering User Satisfaction in Designing Web-Based Portals," Educause Quarterly, 1, 2003. [Online]. Available: https://net.educause.edu/ir/library/pdf/eqm0315.pdf. [Access: January. 2, 2014].

[8] M. Heo, "User Satisfaction with portals: Testing for Factorial validity and Invariance Across Age Groups," Online Information Review, 37(5), 2013, pp. 804-820. [Online]. Available: DOI 10.1108/OIR-06-2012-0099. [Accessed: June. 10, 2014].

[9] P. Koo Yuen, C. K. Ramaiah, "From Picture Press: An Online Exhibition of the SPH Photographs Collection," Journal of Library & Information Technology, May 2013, 33(3), pp. 208-221. [Online]. [Accessed: June. 23, 2014].

[10] S. Kujala, V. Roto, K. V. V. Mattila, E. Karapanos, A. Sinnela, "UX Curve: A method for evaluating long-term user experience," Interacting with Computers, September 2011, 23(5), pp. 473-483. [Online]. Available: DOI: 10.1016/j.intcom.2011.06.005. [Accessed: June. 10, 2014].

[11] Malaysia National Policy of Creative Industry. Ministry of Information, Communication and Culture. 2010. [Online]. Available: http://www.kpkk.gov.my/pdf/dikn.pdf. [Access: April. 3, 2013].

[12] R. Ewen, P. Smith, C. Hulin, C, E. Locke, (1966), "An empirical test of the Herzberg two-factor theory," Journal of Applied Psychology, 1966, 50(6), 544-550.

[13] Doll, W. J. & Torkzadeh, G. (1988). The Measurement of End-User Computing Satisfaction, MIS Quarterly, 12 (2), June 1988, pp. 259-274. [Online]. Available: http://www.jstor.org/stable/248851. [Accessed: June. 10, 2014].

[14] B. Ives, M. H. Olson, J. J. Baroudi, "The measurement of user information satisfaction," Communications of the ACM, 26 (10), October 1983, pp 785-793. [Online]. Available: 10.1145/358413.358430. [Accessed: Aug. 1, 2014].

[15] J. Ang, S. Koh, "Exploring the Relationship Between User Information Satisfaction and Job Satisfaction," International Journal of Information Management, 17(3), June 1997, pp. 169 – 177. [Online]. Available: DOI: 10.1016/S0268-4012(96)00059-X. [Accessed: Aug. 5, 2014].

[16] J. Holland, S. M. Baker, "Customer Participation in Creating Site Brand Loyalty," Journal of Interactive Marketing, 15(4), Autumn (Fall) 2011, pp. 34-35. [Online]. Available: DOI: 10.1002/dir.1021. [Accessed: Aug. 19, 2014].

[17] M. D. Samsur R., S. M. Zabed, Ahmed, "Exploring the Factors Influencing the Usability of Academic Websites: A Case Study in A University Setting," Business Information Review, 30(1), 2013, pp. 40 – 47. [Online]. Available: doi: 10.1177/0266382113482557. [Accessed: Aug. 10, 2014].

[18] Hong-Youl, H., J. Swinder, "An Empirical Test of a Proposed Customer Satisfaction in e-Services," Journal of Services Marketing, 22(5), 2008, pp 399-408. [Online]. Available: DOI 10.1108/08876040810889166. [Accessed: Aug. 12, 2014].

[19] H. Petrie, N. Bevan, "The Evalution of Accessibility, Usability and User Experience," The Universal Access Handbook, C. Stepanidis (ed), CRC Press, 2009. [Online]. Available: www.crcpress.com/product/isbn/9780805862805. [Accessed: Aug. 10, 2013].

[20] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, K. Z. Gajos, "Predicting Users' First Impressions of Website Aesthetics with a Quantification of Perceived Visual Complexity and Colorfulness," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 27 April – 2 May, 2013, Paris, pp. 2049-2058

[21] A. Miniukovich, A.D. Angeli, "Quantification of Interface Visual Complexity," Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces, New York, pp. 153-160.

[22] Azleen I., Mohd Zulkeflee A. B., Rahida A. R. & Mohd Rushdan, Y, "End-User Computing Satisfaction (EUCS) in Computerised Accounting System (CAS): Which the Critical Factors? A Case in Malaysia," Computer and Information Science Journal, 1(2), February 2009, pp. 18 – 24. [Online]. Available: DOI: 10.5539/cis.v2n1p18. [Accessed: Aug. 15, 2014].

[23] J. W. Palmer, "Web Site Usability, Design and Performance Metrics, Journal Information Systems Research, 13 (2), June 2002, pp. 151-167. [Online]. Available: 10.1287/isre.13.2.151.88. [Accessed: Aug. 15, 2014].

[24] J. J. Garrett, "The Elements of User Experience: User-centered Design for the Web," Indiana:New Riders Publications, 2003.

[25] M. T. Lu, W. L. Yeung, "A Framework for Effective Commercial Web Application Development," Internet Research: Electronic Networking Applications and Policy, 1998, 8 (2), pp. 166–173.

[26] N. H. Hashim, J. Murphy, R. Law, "A Review of Hospitality Website Design Frameworks," Information and Communication Technologies in Tourism 2007, pp. 219-230.

[27] H. S. Jung, M. Baker, "Assessing the Market Effectiveness of the World Wide Web in National Tourism Offices," in D. Buhalis, A. M. Tjoa, & J. Jafari (Eds.), Information and Communication Technologies in Tourism 1998, pp. 93–102). New York: Springer-Wien.

[28] C. I. Ho, Y. L. Lee, "The Development of an e-travel Service Quality Scale," Tourism Management, 2007, 28(6), pp. 1434–1449.

[29] J. E. Mills, A. M. Morrison, "Measuring Customer Satisfaction with Online Travel," in A. J. Frew, M. Hitz, & P. O'Connor (Eds.), Information and Communication Technologies in Tourism 2003, pp. 11–28, New York: Springer-Wien.

[30] A. Sambhanthan, A. Good, "Critical Success Factors for Positive User Experience in Hotel Websites: Applying Herzberg's Two Factor Theory for User Experience Modeling", 2013. [Online]. Available: http://arxiv.org/pdf/1302.5196.pdf. [Accessed: May. 15, 2014].

[31] R, Banjbarian, S. Fathi, Z. Rezaei, "Factors Influencing on Customers' E-Satisfaction: A Case Study from Iran," Interdisciplinary Journal of Contemporary Research in Business, January 2012, 3(9), pp. 257-272.

[32] T. L. Childers, C. L. Carr, J. Peck, S. Carson, "Hedonic and Utilitarian Motivations for Online Retail Shopping Behaviour," Journal of Retailing, Winter 2001 77(4), pp. 511 -535. [Online]. Available: DOI: 10.1016/S0022-4359(01)00056-2. [Accessed: May. 15, 2014].

[33] A. V. Hausman, J. S. Siekpe, "The Effect Of Web Interface Features On Consumer Online Purchase Intentions," Journal of Business Research, January 2009, 62(1), pp. 5-13. [Online]. Available: DOI: 10.1016/j.jbusres.2008.01.018. [Accessed: May. 15, 2014].

[34] C. Jones, S. Kim, "Influences Of Retail Brand Trust, Off-Line Patronage, Clothing Involvement And Website Quality On Online Apparel Shopping Intention.," International Journal of Consumer Studies, November 2010, 34(6), pp. 627-637. [Online]. Available: DOI: 10.1111/j.1470-6431.2010.00871.x. [Accessed: May. 15, 2014].

[35] S. Davidson, S. Matheson, "The Evolution of Providing Access to Information: The Fall of the Online Catalog," Legal Reference Services Quarterly, August 3, 2007, 26(57). [Online]. Available: http://ssrn.com/abstract=1337063. [Accessed: Oct. 2, 2014].

[36] W. Hong, J.Y. L. Thong, W. Wong, K. Tam, "Determinants of User Acceptance of Digital Libraries: An Empirical Examination of Individual Differences and System Characteristics," Journal of Management Information Systems, 2002, 18(3), 97-124.

[37] G.Tsakonas, C. Papatheodorou, "Exploring Usefulness And Usability In The Evaluation Of Open Access Digital Libraries," Information Processing & Management, May 2008, 44(3), pp. 1234-1250. [Online]. Available: DOI: 10.1016/j.ipm.2007.07.008. [Accessed: May. 15, 2014].

[38] K. Smith-Yoshimura, C. Shein, "Social Metadata for Libraries, Archives and Museums Part 1: Site Reviews," Dublin, Ohio: OCLC Research. [Online]. Available: http://www.oclc.org/research/publications/library/2011/2011-02.pdf. [Accessed: May. 15, 2014].

[39] "Malaysian Culture and Heritage Digital Bank," 2008-2013. [Online]. Available: http://mchdb.org. [Accessed: Apr. 15, 2013].

[40] "Historypin," July. 2012. [Online]. Available: http://historypin.com. [Accessed: May. 11, 2013].

[41] J. R. Fraenkel, N. E. Wallen, "How to design and evaluate research in education," Seventh Edition. New York: McGraw-Hill, 2010, pp. 7.

[42] R. Hartson, and P. S. Pyla, "The UX Book: Process and Guidelines for Ensuring a Quality User Experience," Waltham: Morgan Kaufmann, 2012, pp. 1 – 36.

Dissecting Genuine and Deceptive Kudos: The Case of Online Hotel Reviews

Snehasish Banerjee, Alton Y. K. Chua

Wee Kim Wee School of Communication & Information
Nanyang Technological University
31 Nanyang Link, SCI Building, Singapore 637718

Abstract—As users continue to rely on online hotel reviews for making purchase decisions, the trend of posting deceptive reviews to heap praises and kudos is gradually becoming a well-established e-business malpractice. Conceivably, it is not trivial for users to distinguish between genuine and deceptive kudos in reviews. Hence, this paper identifies three linguistic cues that could offer telltale signs to distinguish between genuine and deceptive reviews. These linguistic cues include readability, genre and writing style. Drawing data from a publicly available secondary dataset, results indicate that readability and writing style of reviews offer useful clues to distinguish between genuine and deceptive reviews. Specifically, genuine reviews could be more readable and less hyperbolic compared with deceptive entries. With respect to review genre however, the differences were largely blurred. The implications of the findings for theory and practice are highlighted.

Keywords—e-business; user-generated content; online reviews; opinion spam; readability; genre; writing style

I. INTRODUCTION

Users increasingly rely on online reviews for making purchase decisions. In particular, they are often inclined to trust positive reviews, which are meant to applaud products and services, as confirming evidence before making a choice [1]. Furthermore, positive reviews are usually more abundant in review websites than those with either negative or mixed opinions [2]. As a result, it is conceivable that praises and kudos in reviews could significantly impact users' purchase decisions.

However, users need to exercise caution while interpreting positive reviews. Since positive reviews have the potential to boost sales of a given product or service, they offer adequate incentives for organizations to indulge in e-business malpractices such as opinion spamming [3, 4]. For the purpose of this paper, opinion spamming involves posting deceptive reviews containing fictitious praises and kudos with a deliberate attempt to resemble genuine entries. Such a practice is gradually growing into one of the popular e-business tactics among businesses [5, 6, 7]. Hence, consumers could be misled while making purchase decisions.

While it may not be easy to distinguish between genuine and deceptive positive reviews, there could be subtle differences in ways they are written. Hence, this paper seeks to uncover linguistic nuances unique to genuine and deceptive kudos in reviews.

To achieve the objective, it examines authenticity of positive reviews based on three linguistic cues, namely, readability, genre and writing style. Readability refers to the effort and expertise required on the part of users to comprehend the meaning of reviews [7, 8, 9]. Genre refers to the degree to which reviews are informative, which in turn, could influence their distribution of part-of-speech (POS) tags [10, 11, 12]. Writing style refers to authoring approaches such as the use of affective cues, perceptual words and future tense [13, 14, 15, 16].

This paper is significant for both theory and practice. On the theoretical front, it builds on the research areas related to genuine and deceptive reviews through a linguistic analysis. It confirms that nuances in readability and writing style of positive reviews help distinguish between genuine and deceptive entries with reasonable accuracy. In contrast, it suggests that differences between the two could be largely blurred in terms of genre. On the practical front, this paper calls for caution in interpreting reviews, and honesty while posting entries.

The remainder of this paper is structured as follows. The next section presents the related literature. This is followed by the details of the dataset, as well as the operationalization and analysis procedures. The results are presented next. Thereafter, three key findings gleaned from the results are discussed. Finally, the paper concludes with notes on limitations, implications, and directions for future research.

II. LITERATURE REVIEW

The profusion of Web 2.0 has made user-generated content ubiquitous. A specific form of user-generated content that has exponentially grown in popularity and acceptance includes online reviews, which are meant to evaluate products and services. Specifically, reviews for hotels are widely used by users prior to making a booking [17, 18]. They are often perceived as being more genuine and credible vis-à-vis third-party advertisements [19]. Hence, it is no wonder that more than some 80% consumers tend to choose their holiday accommodation based on properties that had been widely applauded in reviews [18].

Users' growing proclivity for hotel reviews thus provides an ideal opportunity for businesses to indulge in opinion spamming. Specifically, positive deceptive reviews heaping praises and kudos are considered priceless. After all, such entries could result in significant financial gains and fames for

businesses [5, 6, 20]. Hence, it is not surprising that posting positive deceptive reviews is fast becoming a well-established e-business malpractice [7].

To aggravate the problem, deceptive reviews are deliberately written to appear genuine. As a result, the lines between them could often be blurred. Nonetheless, drawing from prior studies, this paper argues that even though genuine and deceptive reviews are not easily distinguishable, there could be subtle telltale signs in terms of their readability, genre, and writing style [7, 16, 21]. These three linguistic cues are explained in greater details as follows.

A. Readability

Readability refers to the effort and expertise required on the part of users to comprehend the meaning of reviews [7, 8, 9]. Since genuine and deceptive reviews are written in different contexts, the readability of the two could be different from each other.

Writing genuine reviews is cognitively less challenging than articulating deceptive entries [22]. Moreover, individuals performing a writing task with a high cognitive load tend to write more lucid language than those performing the same task with a lower cognitive load [23]. Hence, deceptive reviews could be more lucid compared with genuine reviews. Stated otherwise, genuine reviews could be less readable vis-à-vis deceptive ones.

However, another school of thought suggests that genuine reviews could be more readable compared with deceptive entries. This is because when users browse reviews, they not only read the entries but also gauge the intelligence of their contributors [24, 25]. Too simplistic reviews might suggest incompetence of the respective contributors in writing sophisticated reviews. Hence, deceptive reviews could be deliberately written using sophisticated language to showcase contributors' competence. This in turn might take a toll on readability.

B. Genre

Genre refers to the degree to which reviews are informative [10, 11, 12]. Writing genuine reviews requires articulating real experiences. On the other hand, writing deceptive reviews requires articulating imaginary experiences that did not occur in reality. Texts written based on real experiences could differ in terms of their genre from accounts based on imagined experiences [26].

There are four genres of text, namely, conversational, task-oriented, informative and imaginative [12]. Among these, genuine reviews could be more informative while deceptive reviews could lean towards being imaginative [11]. Texts of informative genre differ from those of imaginative genre in their distribution of POS tags [10, 11, 12]. Specifically, informative texts contain more adjectives, articles, nouns, and prepositions. In contrast, imaginative texts contain more adverbs, verbs, and pronouns [10, 24].

Among pronouns, personal pronouns in the form of self-references has attracted special attention among the scholarly community. On the one hand, spammers could feel the pangs of conscience while writing deceptive reviews [27]. As a result,

they might use fewer personal pronouns to dissociate themselves from their deceptive comments. On the other hand, spammers could also be enthused by the prospect of deceiving others easily. With great resolve to conceal their deception, it is also possible for them to deliberately enrich deceptive reviews with personal pronouns [16, 28].

C. Writing Style

Writing style refers to authoring approaches used in reviews. For the purpose of this paper, writing style entails the use of affective cues, perceptual words, and future tense [13, 14, 15, 16]. Deceptive reviews could be replete with positive affective cues as a form of exaggeration to create a lasting impact among readers in the online community [29].

Besides, users' physical experiences with hotels are affected by their sensory perceptions [30]. For instance, users' opinion about a hotel could be a function of visual cues such as artwork and aural cues such as music [15]. These cues are reflected in reviews through the use of perceptual words. Conceivably, genuine reviews written after real post-trip experience could be rich in perceptual words.

Additionally, given that positive reviews could favorably impact future sales and revenues of a given hotel [13, 20], deceptive reviews might be articulated not only to describe past experiences in the hotel, but also to express future desires of staying in the same hotel again. Such a writing style might suggest that the positive experiences described in the deceptive reviews are far from being ephemeral. On the other hand, genuine reviews could simply describe past experiences. Hence, they might contain fewer future tense compared with deceptive reviews.

III. METHODS

A. Dataset

A major challenge that hinders research on genuine and deceptive reviews is the difficulty in ascertaining ground truth [31]. After all, it is challenging to validate what is genuine, and what is deceptive in the first place [32]. This has often led scholars to alternatively employ heuristic annotation approaches. For example, [3] deemed duplicate or near duplicate reviews as deceptive ignoring that duplications might at times stem from technical glitches or human errors. Moreover, [33] labeled reviews as either genuine or deceptive with the help of some annotators, who had read a few articles on ways to identify spam. Despite being intuitive, the validity of such heuristic annotation approaches is questionable.

This paper therefore draws ground truth from a publicly available secondary dataset of 800 positive reviews [11]. Specifically, the dataset comprises 400 genuine reviews, and 400 deceptive reviews uniformly distributed across 20 popular hotels in Chicago. Thus, for every hotel, the dataset contained 20 genuine reviews (20 hotels x 20 genuine reviews = 400 genuine reviews altogether), and 20 deceptive reviews (20 hotels x 20 deceptive reviews = 400 deceptive reviews altogether).

This dataset was selected for analysis due to two reasons. First, it is one of the recent works on linguistic differences between genuine and deceptive reviews. Of late, it has been

widely cited by the scholarly community [e.g., 21, 34, 35]. Second, to the best of our knowledge, it is the only publicly available dataset of genuine and deceptive positive reviews till date.

B. Operationalization and Analysis

Readability was operationalized based on four metrics, namely, Gunning-Fog Index (FOG) [36], Coleman-Liau Index (CLI) [37], Automated-Readability Index (ARI) [38], and Flesch-Kincaid Grade Level (FKG) [39, 40]. Each metric employs a set of unique constants and depend on factors such as number of characters per word, number of words per sentence, and number of syllables per word. More detailed description about these metrics can be found in works such as [8], [9] and [41]. A Java program was written to compute these metrics. Lower values for the readability metrics suggest greater readability. Among the four metrics, FOG and CLI specifically indicate complexity, while ARI and FKG are proxies for reading difficulty [8]. Therefore, FOG and CLI scores for every review were averaged to create a composite index for complexity. Likewise, ARI and FKG scores for every review were averaged to create a composite index for reading difficulty. Finally, review readability was measured in terms of two indicators, namely, (1) complexity, and (2) reading difficulty.

Review genre was operationalized on the basis of the POS tag distributions in reviews. Specifically, the following eight POS tags were considered: (1) adjective, (2) article, (3) noun, (4) preposition, (5) adverb, (6) verb, (7) pronoun, and (8) personal pronoun. While the first four are expected to be higher in genuine reviews, the next four could be higher in deceptive reviews [10, 12, 16]. The fractions of each of these POS tags in reviews were computed using Stanford Parser's POS tagger [42].

Review writing style was operationalized as the proportion of (1) positive cues, (2) perceptual words, and (3) future tense used in reviews. These indicators were measured using the Linguistic Inquiry and Word Count (LIWC) software [43]. It is an automated text analysis tool that offers reliable dictionaries to compute such linguistic indicators.

To sum up, this paper includes a total of 13 independent variables (IVs) for analysis as follows: the two readability indicators, the eight POS tags, and the three writing style indicators. The categorical dependent variable (DV) comprises review authenticity. Since this paper seeks to examine review authenticity as a function of readability, genre and writing style, the DV was dummy-coded such that 1 indicates genuine reviews and 0 denotes deceptive reviews. Given its dichotomous nature, binomial logistic regression was used for data analysis [44]. The coefficients of logistic regression estimate the odds ratio, indicating the extent to which the IVs in the model could predict review authenticity.

To diagnose potential problems of multicollinearity in the logistic regression model, the variance inflation factors (VIF) for all the 13 IVs were examined. The VIF values were found to be less than 10, suggesting that multicollinearity did not exist [45]. Another potential problem of logistic regression is the presence of outliers in the solution [46]. In particular, cases with standardized residual values of above 2.5 or below -2.5 could be problematic [47]. Only one out of 800 reviews was found to have exceeded the acceptable threshold, and was retained for the analysis.

After analysis, the performance of the logistic regression model was probed using Omnibus test, and Hosmer-Lemeshow goodness-of-fit test. The extent to which the model could account for the variability in the dependent variable was examined using two pseudo-R^2 measures, namely, Cox and Snell R^2, as well as Nagelkerke R^2 [47]. Finally, the ability of the model to differentiate between genuine and deceptive reviews was checked using 10-fold cross-validation. This facilitates checking the model's stability in distinguishing between genuine and deceptive reviews for unknown datasets.

IV. RESULTS

Table I presents the non-parametric inter-correlations among the variables involved in the analysis. Variables 1 through 13 represent the 13 IVs, while variable 14 comprises the DV, namely, review authenticity. With respect to the DV, 10 of the 13 IVs had statistically significant correlations.

TABLE.I. SPEARMAN NON-PARAMETRIC INTER-CORRELATIONS BETWEEN VARIABLES INVOLVED IN THE ANALYSIS

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1 Complexity	1													
2 Reading diff.	.92*	1												
3 Adjective	.09*	.02	1											
4 Article	-.01	.01	-.02	1										
5 Noun	.14*	.07	.02	.07*	1									
6 Preposition	.15*	.20*	-.27*	-.05	.06	1								
7 Adverb	-.20*	-.19*	.06	-.19*	-.33*	-.26*	1							
8 Verb	-.34*	-.31*	-.10*	-.11*	-.50*	-.31*	.26*	1						
9 Pronoun	-.16*	-.08*	-.40*	-.31*	-.48*	-.04	.12*	.39*	1					
10 Pers. pronoun	-.13*	-.07*	-.42*	-.28*	-.41*	.01	.11*	.33*	.86*	1				
11 Positive cues	.09*	-.01	.39*	-.03*	-.02	-.29*	.14*	-.02	-.13*	-.14*	1			
12 Percep. words	.05	.02	.10*	-.09*	-.01	-.01	-.03	-.02	.02	.02	.11*	1		
13 Future tense	-.01	-.01	-.05	-.12*	-.18*	-.12*	.24*	.17*	.14*	.12*	.04	.01	1	
14 DV	-.19*	-.18*	.12*	.06	.15*	.02	-.03	-.10*	-.31*	-.33*	-.11*	-.11*	-.14*	1

* $p < 0.05$

Descriptive statistics of the dataset are presented in Table II and Table III. In particular, Table II provides mean, standard deviation, minimum and maximum of the 13 IVs for the full dataset of 800 reviews (full), the subset comprising 400 genuine reviews (genu), and the subset containing 400 deceptive reviews (decep). Thereafter, range as well as first, second and third quartiles of the IVs are presented in Table III.

TABLE.II. MEAN, STANDARD DEVIATION, MINIMUM AND MAXIMUM OF THE 13 IVS

IVs	Mean			Standard Deviation			Minimum			Maximum		
	full	genu	decep	full	genu	decep	full	genu	decep	full	genu	decep
Complexity	9.26	8.94	9.59	2.42	2.67	2.09	3	3	4	27.5	27.5	18.5
Reading diff.	6.78	6.53	7.04	3.62	4.39	2.63	.5	.5	1.5	45.5	45.5	23.5
Adjective	10.48	10.88	10.09	3.33	3.47	3.14	2	2	3	27	27	23
Article	10.10	10.26	9.95	2.75	2.90	2.57	0	0	2	18	18	18
Noun	26.90	27.67	26.13	4.97	5.18	4.64	15	15	15	44	44	43
Preposition	12.13	12.19	12.07	3.02	2.93	3.12	2	3	2	20	20	20
Adverb	5.11	5.00	5.23	2.60	2.36	2.82	0	0	0	18	15	18
Verb	12.24	11.87	12.60	3.24	3.17	3.27	2	2	3	24	20	24
Pronoun	10.59	9.38	11.80	3.96	3.63	3.91	0	0	1	23	21	23
Pers. pronoun	7.04	5.96	8.12	3.31	2.92	3.33	0	0	0	17	14	17
Positive cues	6.89	6.63	7.15	2.95	2.99	2.89	0	0	1	19	19	19
Percep. words	2.08	1.90	2.26	1.59	1.52	1.65	0	0	0	9	9	8
Future tense	0.84	0.71	0.98	0.87	0.77	0.95	0	0	0	5	4	5

TABLE.III. RANGE AND QUARTILES OF THE 13 IVS

IVs	Range			First Quartile			Second Quartile			Third Quartile		
	full	genu	decep	full	genu	decep	full	genu	decep	full	genu	decep
Complexity	24.5	24.5	14.5	8	7.5	8	9	8.5	9.5	10.5	10	11
Reading diff.	45	45	22	5	4.5	5.5	6.5	6	6.5	8	7.5	8.5
Adjective	25	25	20	8	9	8	10	10	10	12	13	12
Article	18	18	16	8	8.3	8	10	10	10	12	12	12
Noun	29	29	28	24	24	23	26	27	26	30	31	29
Preposition	18	17	18	10	10	10	12	12	12	14	14	14
Adverb	18	15	18	3	3	3	5	5	5	7	6	7
Verb	22	18	21	10	10	10	12	12	12	14	14	15
Pronoun	23	21	22	7	7	9	11	9	12	13	12	14
Pers. pronoun	17	14	17	4	4	6	7	6	8	9	8	11
Positive cues	19	19	18	5	5	5	6	6	7	8	8	8.75
Percep. words	9	9	8	1	1	1	2	2	2	3	3	3
Future tense	5	4	5	0	0	0	1	1	1	1	1	1

For the logistic regression model, result of the Omnibus test indicates acceptable performance of the model (χ^2 = 206.74; df = 13; - 2 log likelihood = 902.29; p < 0.001). The Hosmer-Lemeshow goodness-of-fit test indicated a non-significant result (χ^2 = 7.11; df = 8; p = 0.53), which suggests that the model fits well with the data. Cox and Snell R^2 was 0.27 while Nagelkerke R^2 was 0.36. Thus, around 27% to 36% of the variability in review authenticity could be explained by the model. Using 10-fold cross-validation, the model accurately predicted 281 of the 400 genuine reviews, and hence had a genuine review prediction accuracy of 70.25%. On the other hand, it could accurately predict 283 of the 400 deceptive reviews, and hence had a deceptive review prediction accuracy of 70.75%. Overall, it recorded an accuracy of 70.50%.

Results further indicate that the two readability indicators, namely, complexity and difficulty could significantly predict if reviews were genuine or deceptive. In particular, complexity was negatively related to review authenticity [β = -0.75, Exp(β) = 0.47, p < 0.001]. The higher the value of complexity for a given review, the lower was its likelihood to be genuine. Put differently, genuine reviews had lower values for complexity compared with deceptive reviews, suggesting that the former is linguistically less complex than the latter. However, reading difficulty was positively related to review authenticity [β = 0.37, Exp(β) = 1.45, p < 0.001]. In other words, genuine were generally more difficult to be read compared with deceptive entries.

The discordant finding between complexity and reading difficulty could be vestige of the uniqueness of the four readability metrics. Furthermore, it supports the argument that the two readability indicators, namely, complexity and reading difficulty, do not necessarily imply each other [8]. However, to the best of our knowledge, not much research hitherto has disinterred the nuances between complexity and reading difficulty in the context of genuine and deceptive reviews. To further tease out nuances, the factors that affect FOG, CLI, ARI and FKG were delved deeper.

These four readability metrics are primarily affected by three constituent factors, namely, (1) average characters per word, (2) average words per sentence, and (3) average syllables per word [8, 9, 41]. To disinter variations between genuine and deceptive reviews based on the three factors, independent samples t-tests were performed. In terms of average characters per word, genuine reviews (M = 4.40, SD = 0.28) did not significantly differ from deceptive reviews (M = 4.40, SD = 0.30). With respect to average words per sentence too, there was no significant difference between genuine (M = 14.46, SD = 9.72) and deceptive reviews (M = 15.07, SD = 5.26). However in terms of average syllables per word, there was a significant difference between genuine reviews (M = 1.40, SD = 0.10) and deceptive reviews (M = 1.44, SD = 0.12) [t(779.85) = 4.53, p < 0.001]. Given that genuine reviews used significantly lower number of syllables per word compared with deceptive reviews, the former seems to fare better in terms of readability.

Among the eight POS tags, articles, pronouns and personal pronouns turned out to be significant predictors of review authenticity. All three were negatively related to the DV as

follows: articles [β = -0.12, Exp(β) = 0.89, p < 0.01], pronouns [β = -0.11, Exp(β) = 0.89, p < 0.05], and personal pronouns [β = -0.19, Exp(β) = 0.82, p < 0.001]. In other words, reviews with fewer articles, pronouns and personal pronouns were more likely to be genuine. On the other hand, reviews that comprised more articles, pronouns and personal pronouns were more likely to be deceptive.

With respect to writing style, all the three metrics, namely, the use of positive cues, perceptual words and future tense, emerged as significant predictors of review authenticity. All three were negatively related to the DV as follows: positive cues [β = -0.10, Exp(β) = 0.90, p < 0.01], perceptual words [β = -0.15, Exp(β) = 0.86, p < 0.01], and future tense [β = -0.30, Exp(β) = 0.74, p < 0.01]. Thus, it appears that deceptive reviews were generally more richly embellished with positive cues, perceptual words and future tense compared with genuine reviews. Table IV summarizes the extent to which the 13 IVs in the model could predict review authenticity.

TABLE.IV. RESULTS OF THE LOGISTIC REGRESSION MODEL

Linguistic Cues	IVs	β	SE	Wald	Exp(β)
Readability	Complexity[***]	-0.75	0.10	52.16	0.47
	Reading difficulty[***]	0.37	0.07	31.87	1.45
Genre	Adjective	0.02	0.04	0.50	1.02
	Article[**]	-0.12	0.04	10.69	0.89
	Noun	-0.02	0.02	0.39	0.98
	Preposition	-0.04	0.03	1.25	0.96
	Adverb	-0.04	0.04	1.31	0.95
	Verb	-0.06	0.03	2.67	0.95
	Pronoun[*]	-0.11	0.04	6.50	0.89
	Personal pronoun[***]	-0.19	0.05	15.01	0.82
Writing Style	Positive cues[**]	-0.10	0.03	9.52	0.90
	Perceptual words[**]	-0.15	0.05	7.54	0.86
	Future tense[**]	-0.30	0.10	8.52	0.74
Pseudo-R^2	0.27 (Cox and Snell), 0.36 (Nagelkerke)				
Accuracy	70.50% (10-fold cross-validation)				

[***] P < 0.001, [**] P < 0.01, [*] P < 0.05

V. DISCUSSION

Three key findings could be gleaned from the results. First, in terms of readability, genuine reviews were more readable than deceptive reviews. For example, a readable genuine review in the dataset indicated, *"...the hotel is centrally located...is less than a block away. perfect location! the suites are huge with comfy beds...also they have a free dinner..."* In contrast, a less readable deceptive review stated, *"...boasts a modern fitness center that feature free weights, a cardio room, dry saunas, as well as, masseurs...elegant with a touch of historic decor...breath-taking view of Chicago, as well as, had an in-room mini-bar, terry-cloth bath robes, over-sized desks, high-speed internet access, and a 37-inch Hi Def LCD Television..."* This finding is consistent with prior research which suggested that deceptive content could be less readable than the genuine counterpart [21, 48, 49]. Unlike users articulating genuine experiences, spammers were overly ostentatious in reflecting their competence in writing sophisticated reviews [24, 25]. That could be why deceptive reviews comprised significantly higher syllables per word as compared with genuine entries. Interpreting this finding on the basis of self-presentation effect [50], users writing genuine

reviews appear less motivated than spammers to make ostentation of linguistic competence.

Second, in terms of genre, genuine and deceptive reviews appeared to share similar levels of informativeness. For example, an informative genuine review pointed, "*…large room with 2 double beds and 2 bathrooms, The TV was Ok, a 27' CRT Flat Screen… The breakfast is charged, 20 dollars…close to metro station…*" Likewise, an informative deceptive review expressed, "*…located in the heart of…has a 24 hour business center providing high-speed internet access, fax, and photocopying services…in-room mini-bar…and a 37-inch Hi Def LCD Television…*" Prior research suggests that genuine and deceptive reviews could be informative and imaginative respectively [11, 12]. While the former could contain more adjectives, articles, nouns and prepositions, the latter could be richer in adverbs, verbs, pronouns and personal pronouns. However, only articles, pronouns and personal pronouns could significantly distinguish between genuine and deceptive reviews. Furthermore, the finding that reviews with fewer articles were more likely to be genuine contradicts literature on text genre [10, 12]. The dominance of personal pronouns in deceptive reviews over genuine entries reflects the lack of guilt among spammers. Although prior research expects them to feel guilty and use less self-references to dissociate themselves from deceptive content [26, 27, 51], such a phenomenon was generally inconspicuous. A deceptive review rich in personal pronouns stated, "*…I came with very little…my deluxe room supplied me with everything that I needed…I will be back…*" This suggests that spammers could be adept enough to blur the lines between genuine and deceptive reviews with respect to text genre.

Third, in terms of writing style, genuine reviews appeared less hyperbolic compared with deceptive ones. Consistent with extant literature [16, 29], deceptive reviews seemed to include significantly more positive affective cues than genuine reviews. For example, a deceptive review pointed, "*The hotel was very nice; Service was great, everyone was very friendly. The room was very elegant…had a great experience…pleasant staff…I left here well rested and happy.*" In order to steer users' impression on hotels towards a positive light, deceptive reviews seemed to contain more perceptual words than genuine entries. This was perhaps deliberately done to appeal to the sensory perceptions of the online community [15, 30]. A highly perceptual deceptive review illustrated, "*…view from my windows was stunning, as I looked out I could see the beautiful…room had a nice airy feel but was also warm…bed was very comfortable…*" The excessive use of future tense in deceptive reviews might have been used to assure that the positive experiences at the hotel were not ephemeral. After all, positive reviews could be highly influential in stimulating future sales and revenues of a given hotel [13, 20]. A deceptive review rich in future tense expressed, "*…will leave you absolutely relaxing…will leave you wanting to visit the moment you leave.*"

VI. Conclusion

As users continue to rely on online hotel reviews for making purchase decisions, the trend of posting deceptive reviews to heap praises and kudos is gradually becoming a well-established e-business malpractice. Hence, this paper attempted to distinguish between genuine and deceptive reviews using linguistic analysis. In particular, it investigated the extent to which linguistic differences between genuine and deceptive reviews in terms of readability, genre, and writing style could predict review authenticity. Drawing data from a publicly available secondary dataset, results indicate that readability and writing style of reviews are useful clues to distinguish between genuine and deceptive reviews. Specifically, genuine reviews could be more readable and less hyperbolic compared with deceptive entries. With respect to review genre however, the differences were largely blurred.

It should be acknowledged that the findings of the paper are somewhat constrained by the dataset used for analysis. For one, it comprised only positive reviews. Even though it facilitated distinguishing between genuine and deceptive kudos in reviews, the findings are not generalizable to negative reviews that are meant to criticize hotels, or mixed reviews that highlight both merits and demerits of hotels. Moreover, the reviews were meant for some popular hotels in Chicago. Hence, it is unknown if the findings could be extrapolated to all types of hotels located in various geographical locations. Moreover, the dataset size of 800 reviews was not overly large. This could have resulted in inaccurate findings. Nonetheless, this paper does offer implications for both theory and practice.

On the theoretical front, this paper augments prior studies such as [3, 7, 11] by conducting a linguistic analysis of genuine and deceptive reviews. It demonstrates that readability and writing style of reviews could significantly distinguish between genuine and deceptive reviews. While genuine reviews could be more readable vis-à-vis deceptive entries, the former could be articulated with a less hyperbolic writing style. In terms of genre however, this paper demonstrates that genuine and deceptive reviews are equally informative. This finding is at odds with prior studies such as [52] and [26], which expected genuine reviews to be more informative than the deceptive counterpart.

On the practical front, this paper serves as an eye-opener pointing that all positive reviews should not be trusted. Prior research suggests that when users browse reviews, they could be tempted to trust positive entries that are generally abundant [1, 2, 53]. However, not every positive review is necessarily authentic [7, 11]. Hence, users need to exercise caution. They could lean on the findings of this paper to conjecture which reviews are likely to be genuine accounts of post-trip experiences and hence, can be relied for travel planning. Based on the findings, moderators of review websites could automatically recommend reviews that are potentially genuine and flag off those that are likely to be deceptive. The findings can thus play a significant role in preventing users from being victims of deceptive opinion spamming. This will aid more informed travel planning, thereby mitigating hotels' e-business malpractice of promoting themselves through deceptive kudos in reviews.

This paper further offers a few potential directions for future research. For one, the dataset could be expanded beyond reviews for hotels to include those for other products and services such as consumer electronics, or downloadable

applications. Another possible direction could include analysis of the extent to which linguistic differences predict review authenticity across positive, negative and mixed reviews. Such studies could help extend the theoretical boundaries of this paper. Moreover, the finding that genuine and deceptive reviews are equally informative is significant for further research. Spammers are increasingly becoming smarter to blur the lines between the two. As they learn the patterns to mimic genuine reviews, it is important for the scholarly community to catch up. Perhaps in due course of time, linguistic analysis alone might no longer be sufficient to distinguish between genuine and deceptive reviews. In this vein, this paper serves to pique further scholarly inquiry into this research theme from aspects beyond linguistic analysis.

REFERENCES

[1] Y. Pan, and J. Q. Zhang, "Born unequal: A study of the helpfulness of user-generated product reviews," Journal of Retailing, vol. 87, 2011, pp. 598-612.

[2] N. Hu, J. Zhang, and P. A. Pavlou, "Overcoming the J-shaped distribution of product reviews," Communications of the ACM, vol. 52, 2009, pp. 144-147.

[3] N. Jindal, and B. Liu, "Opinion spam and analysis," Proceedings of the International Conference on Web Search and Web Data Mining, 2008, pp. 219-230.

[4] D. Fallis, D. "A conceptual analysis of disinformation," Proceedings of the iConference, 2009, Chapel Hill, NC. Available at http://ideals.illinois.edu/bitstream/handle/2142/15205/fallis_disinfo1.pdf

[5] E. Allen, "Dear staff. We could do with some positive comments: Hotel boss is caught telling his workers to post fake reviews on TripAdvisor," MailOnline (2012, October 9). Internet: http://www.dailymail.co.uk/new s/article-2214974/Hotel-boss-caught-telling-workers-post-fake-reviews-TripAdvisor.html [Dec. 26, 2013].

[6] C. Fernandez, "Tripadvisor bribes: Hotel owners offer free rooms in return for glowing reviews," MailOnline (2011, July 11). Internet: http://www.dailymail.co.uk/travel/article-2013391/Tripadvisor-Hotel-owners-bribe-guests-return-good-reviews.html [Jun. 5, 2013].

[7] N. Hu, I. Bose, N. S. Koh, and L. Liu, "Manipulation of online reviews: An analysis of ratings, readability, and sentiments," Decision Support Systems, vol. 52, 2012, pp. 674-684.

[8] N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso, "Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content," Electronic Commerce Research and Applications, vol. 11, 2012, pp. 205-217.

[9] B. L. Zakaluk, and S. J. Samuels, Readability: Its past, present and future. Newark: International Reading Association, 1988.

[10] G. Leech, P. Rayson, and A. Wilson, Word frequencies in written and spoken English: Based on the British National Corpus. London: Longman, 2001.

[11] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," Proceedings of the Association for Computational Linguistics, 2011, pp. 309-319.

[12] P. Rayson, A. Wilson, and G. Leech, "Grammatical word class variation within the British National Corpus sampler," Language and Computers, vol. 36, 2001, pp. 295-306.

[13] W. Duan, B. Gu, and A. B. Whinston, "Do online reviews matter? An empirical investigation of panel data," Decision Support Systems, vol. 45, 2008, pp. 1007-1016.

[14] H. C. Lench, S. A. Flores, and S. W. Bench, "Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations," Psychological Bulletin, vol. 137, 2011, pp. 834-855.

[15] I. Y. Lin, "Evaluating a servicescape: The effect of cognition and emotion," International Journal of Hospitality Management, vol. 23, 2004, pp. 163-178.

[16] K. H. Yoo, and U. Gretzel, "Comparison of deceptive and truthful travel reviews," ICT in Tourism, 2009, pp. 37-47.

[17] J. Fotis, D. Buhalis, and N. Rossides, "Social media impact on holiday travel planning: The case of the Russian and the FSU markets," International Journal of Online Marketing, vol. 1, 2011, pp. 1-19.

[18] K. L. Sidali, H. Schulze, and A. Spiller, "The impact of online reviews on the choice of holiday accommodations," ICT in Tourism, 2009, pp. 87-98.

[19] S. Senecal, and J. Nantel, "The influence of online product recommendations on consumers' online choices," Journal of Retailing, vol. 80, 2004, pp. 159-169.

[20] L. S. Chen, and J. Y. Lin, "A study on review manipulation classification using decision tree," International Conference on Service Systems and Service Management, 2013, pp. 680-685.

[21] T. Ong, M. Mannino, and D. Gregg, "Linguistic characteristics of shill reviews," Electronic Commerce Research and Applications, vol. 13, 2014, pp. 69-78.

[22] M. Zuckerman, B. M. DePaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception," Advances in Experimental Social Psychology, vol. 14, 1981, pp. 1-59.

[23] J. K. Burgoon, and T. Qin, "The dynamic nature of deceptive verbal communication," Journal of Language and Social Psychology, vol. 25, 2006, pp. 76-96.

[24] L. Connors, S. M. Mudambi, and D. Schuff, "Is it the review or the reviewer? A multi-method approach to determine the antecedents of online review helpfulness," Proceedings of the IEEE Hawaii International Conference on System Sciences, 2011, pp. 1-10.

[25] P. F. Wu, H. van der Heijden, and N. Korfiatis, "The influences of negativity and review quality on the helpfulness of online reviews," Proceedings of the International Conference on Information Systems, 2011, pp. 1-10.

[26] A. Vrij, K. Edward, K. P. Roberts, and R. Bull, "Detecting deceit via analysis of verbal and nonverbal behavior," Journal of Nonverbal Behavior, vol. 24, 2000, pp. 239-264.

[27] A. Vartapetiance, L. Gillam, "'I don't know where he is not': Does deception research yet offer a basis for deception detectives?," Proceedings of Computational Approaches to Deception Detection, 2012, pp. 5-14.

[28] S. Godin, All Marketers Are Liars. New York: Portfolio, 2005.

[29] C. Maurer, and S. Schaich, "Online customer reviews used as complaint management tool," ICT in Tourism, 2011, pp. 499-511.

[30] H. R.Schiffman, Sensation and Perception, 5th Edition. New York: Wiley, 2001.

[31] S. Gokhman, J. Hancock, P. Prabhu, M. Ott, and C. Cardie, "In search of a gold standard in studies of deception," Proceedings of the Workshop on Computational Approaches to Deception Detection, 2012, pp. 23-30.

[32] G. Wu, D. Greene, and P. Cunningham, "Merging multiple criteria to identify suspicious reviews," Proceedings of the ACM Conference on Recommender systems, 2010, pp. 241-244.

[33] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," Proceedings of the International Joint Conference on Artificial Intelligence, 2011, pp. 2488-2493.

[34] J. Parapar, D. E. Losada, and A. Barreiro, "Combining psycho-linguistic, content-based and chat-based features to detect predation in chatrooms," Journal of Universal Computer Science, vol. 20, 2014, pp. 213-239.

[35] H. Sun, A. Morales, and X. Yan, "Synthetic review spamming and defense," Proceedings of the International Conference on Knowledge Discovery and Data Mining, 2013, pp. 1088-1096.

[36] R. Gunning, The technique of clear writing. New York: McGraw-Hill, 1952.

[37] M. Coleman, and T. L. Liau, "A computer readability formula designed for machine scoring," Journal of Applied Psychology, vol. 60, 1975, pp. 283-284.

[38] E. A. Smith, and R. J. Senter, Automated readability index, AMRL-TR-66-22. Wright-Patterson AFB, Ohio: Aerospace Medical Division, 1967.

[39] R. F. Flesch, How to test readability. New York: Harper, 1951.

[40] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, No. RBR-8-75, Naval Technical Training Command, Millington TN Research Branch, 1975.

[41] W. H. DuBay, The Principles of Readability, 2004. Internet: http://files.eric.ed.gov/fulltext/ED490073.pdf [Sep. 30, 2014].

[42] D. Klein, and C. D. Manning, "Accurate unlexicalized parsing," Proceedings of the Association for Computational Linguistics, 2003, pp. 423-430.

[43] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of LIWC 2007," Austin, TX: LIWC.net, 2007.

[44] J. F. Hair, R. E. Anderson, R. L. Tatham, and W. C. Black, Multivariate data analysis, 5th edition. New Jersey: Prentice Hall, 1998.

[45] R. Myers, Classical and Modern Regression with Applications. Boston: Duxbury Press, 1990.

[46] B. G. Tabachnick, and L. S. Fidell, Using multivariate statistics, 3rd Edition. New York: Harper Collins, 1996.

[47] J. Pallant, SPSS survival manual: A step by step guide to data analysus using SPSS for Windows (version 12). Sydney: Allen & Unwin, 2005.

[48] R. L. Daft, and R. H. Lengel, "Information richness: A new approach to managerial behavior and organizational design," in Research in organizational behavior, L. L. Cummings, and B.M. Staw, Eds. Homewood, Illinois: JAI Press, 1984, pp. 191-233.

[49] K. Moffitt, and M. B. Burns, "What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports," Proceedings of Americas Conference on Information Systems, 2009, pp. 399:1-10.

[50] M. R. Leary, Self-presentation: Impression management and interpersonal behavior. Wisconsin: Brown & Benchmark, 1995.

[51] A. Vrij, K. Edward, and R. Bull, "Stereotypical verbal and nonverbal responses while deceiving others," Personality and Social Psychology Bulletin, vol. 27, 2001, pp. 899-909.

[52] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," Personality and Social Psychology Bulletin, vol. 29, 2003, pp. 665-675.

[53] P. A. Pavlou, and A. Dimoka, "The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation," Information Systems Research, vol. 17, 2006, pp. 392-414.

Process Knowledge Discovery in Social BPM

Mohammad Ehson Rangiha

School of Engineering Computer Science and
Mathematics
City University London,
UK

Bill Karakostas

School of Engineering Computer Science and
Mathematics
City University London,
UK

Abstract—the utilization of process knowledge for future execution is an effective way of improving the efficiency of business processes and benefit from the knowledge captured in previous executions. This paper attempts to discuss how social tagging can be used in the context of social business process management to assist and support the execution of business processes in a social environment. We believe such an approach is a step forward towards producing a comprehensive model for social business process management.

Keywords—Social BPM; BPM; Goal-Based Modelling; Process Recommendation; Social Tagging

I. Introduction

Business Process Management (BPM) is a systematic approach which provides a platform to take full control of the business processes from the beginning until the end. This is achieved through the management, measurement and the improvement of the company processes [2].

With the increase in the usage of social software and user collaboration in the different areas of day to day life, recently traditional BPM models have been influenced by social characteristics which promise to improve and overcome the limitations of traditional BPM approaches. Initial investigations in the area of social BPM has taken place, however these research are still in the initial stages [3]. Social BPM is the intersection between social software and BPM and the integration of social elements into the different stages of the BPM lifecycle.

Goal-based modelling has been proposed as a guiding theory to support social BPM overcome the sequential nature of traditional BPM systems and also enable flexibility during the design and execution of the processes [30]. In this approach goals are defined and set as the intended outcomes to be achieved [6].

The proposed SBPM framework which is inspired by goal-based modelling, does not enforce tasks during the execution of the processes, this has brought about the idea of 'process recommender'. Furthermore process ownership in our SBPM model does not follow a top-down imposing mechanism where the process owners are selected by a higher authority, rather it follows a 'role assignment' approach which is informed by the social behavior of the user and his/her profile static and feedback data [6].

This paper aims to unfold the key elements of the process recommender mechanism in the SBPM model and is structured in the following way:

In section 2, the theories used in the formation of the proposed SBPM approach are presented, explaining how they have been adapted in the context of social BPM. This is followed by a high-level model of SBPM illustrating its main components in section 3 in order to position out research at hand accurately.

Section 4 explains the task recommender mechanism, illustrating how it works and presenting a meta-model for this components and introducing its main elements. Furthermore to clarify what has been discussed, an example applied to the theories presented is demonstrated to further clarify the function of the take recommender. Finally, section 5 presents a summary of the research, explains some of the contributions of this approach of SBPM to the field and after identifying some of the limitations of the presented approach, lays the ground for future research and investigation.

II. Goal-Based Social Business Process Management (SBPM)

A. Overview of BPM

Business Process Management is a discipline which focuses on optimizing, managing and running of business processes [28] and according to Aalst et al [12] it consists of the following main stages: process design/discovery, configuration/implementation, enactment/execution and evaluation.

Limitations such as Lack of Information Fusion, Model-Reality Divide, Information Pass-On Threshold and Lost Innovation and Strict Access-Control [4], have initiated new research such as integrating social characteristics into the BPM lifecycle.

B. Overview of SBPM

Social Business Process Management is an approach for engaging a range of users in the design, implementation, enactment and evaluation of business processes in order to improve the BPM lifecycle [27]. SBPM is the intersection of BPM with social software to optimize the efficiency of traditional BPM systems [26], and processes are designed and enhanced by adapting various social technologies [8]. Gartner discusses the collaborative nature of social BPM stating that it is a concept which describes collaboratively designed and iterated processes [9]. There has been some research in this area; however none of these have produced a formalized framework yet [3] [5].

SBPM aims to account for the unplanned participation of the users and alteration of the business processes to attain a

certain level of flexibility and complete enactment of the process steps more efficiently [1]. This is achieved through allowing the users to be part of the design of the processes, therefore instead of a group of analysts designing the processes and passing it onto the end users, the users themselves are also involved in the design of the processes this eliminates the necessity for the presence of a group of process designer [8].

In the current model of SBPM, the analysts and developers are using social software to improve their experience of BPM [8] [26], however social elements have not been fully incorporated throughout the BPM lifecycle and specifically during runtime of the processes. Only recently, [6][30] [26] have proposed goal-based modeling which has been a step forward in formalizing a SBPM framework.

A number of benefits for SBPM has been identified as discussed in [5] such as: exploitation of weak ties and implicit knowledge [1][3], transparency [3][5], decision distribution [3] [16], knowledge sharing [3][16][5]. At the same time, a number of potential limitations and disadvantages have also been mentioned about SBPM, these include: learning effort [4][1], security [4][5], quality [4], difficulty to evaluate [5], process management [26].

C. Goal-Based Modelling

Goal- based modelling [31] is the approach which has been used [30] in order to overcome the rigid and sequential nature of traditional BPM systems since flexibility during runtime of the process cannot be achieved with sequential workflows [5]. Goal-based modelling provides a flexible process-flow whilst keeping the integrity of the processes intact in the social BPM framework. Although this approach has been used extensively in requirements engineering [23] [22], however it has had not been adapted in the context of SBPM.

Goals are states which are reached through the execution of a number of steps [28] and they provide the intention behind the activities which need to be performed [30]. The type of goals mentioned here are process goals which are set by the different process areas of an organization and upon completion; their status is changed to 'satisfied'. These goals may well have dependencies and other sub-goals which need to be achieved first before the main goal can be satisfied [29], all of these are modelled in goal-based business process approach.

The goals which are set by the organization are left to the specific user(s) to decide how they would like to achieve them [1] [24]. In other words goal-based modelling allows the

capturing of the 'what' without specifying the 'how' as [28] puts it. The 'how' is similar to the *imperative approach* in which the tasks order and sequence is explicitly defined, and the 'what' is similar to the *declarative approach* in which the end goal is important and not how it is achieved [68]. These two approaches have been presented in the following figure:

Depending on how detailed the goals are set; the flexibility given to the user to carry out the relevant tasks also varies. Normally what would happen is that after a few iteration of completing a goal, the approach and series of steps which needs to be taken gradually becomes more structured [5]. These goals can then be re-evaluated going forward based on user feedback and experience of running the processes.

In our proposed approach for processes recommendation, the predefined goals are recommended to the user and the user has a degree of flexibility to achieve the proposed goal. Unlike traditional BPM systems, in this approach sequence of actions are not enforced upon the user, rather it only guides the user throughout the enactment of the processes by enabling collaboration and participation from the user community [5].

Giving the user the flexibility to decide how he/she would like to attain a goal would only be feasible in processes where the process flow is not programmed and hard-coded into the system.

III. OVERALL STRUCTURE OF SBPM

[7] Presents the architecture of the proposed SBPM framework and outlines the main components of it, namely the task recommender system and role-assignment mechanism. Fig 1presents a high-level view of the various elements in the proposed social BPM model and illustrates how the two main components of our approach is linked with the rest of the environment. Here some of the important parts of this model is explained:

Process Goals: if the overall goal of the process which the users are trying to achieve.

Tasks: each goal have a number of tasks which need to be fulfilled in order for the overall goal to be satisfied. The number or complexity of the tasks are dependent on the nature of the domain specific processes.

Database: static, behavioral and feedback data related to the users and the discussions are all stored in the database and used for recommending tasks and roles as has been presented in [7]

Social User Goals : these are the reasons and motives behind different users involvement and participation in the business processes.
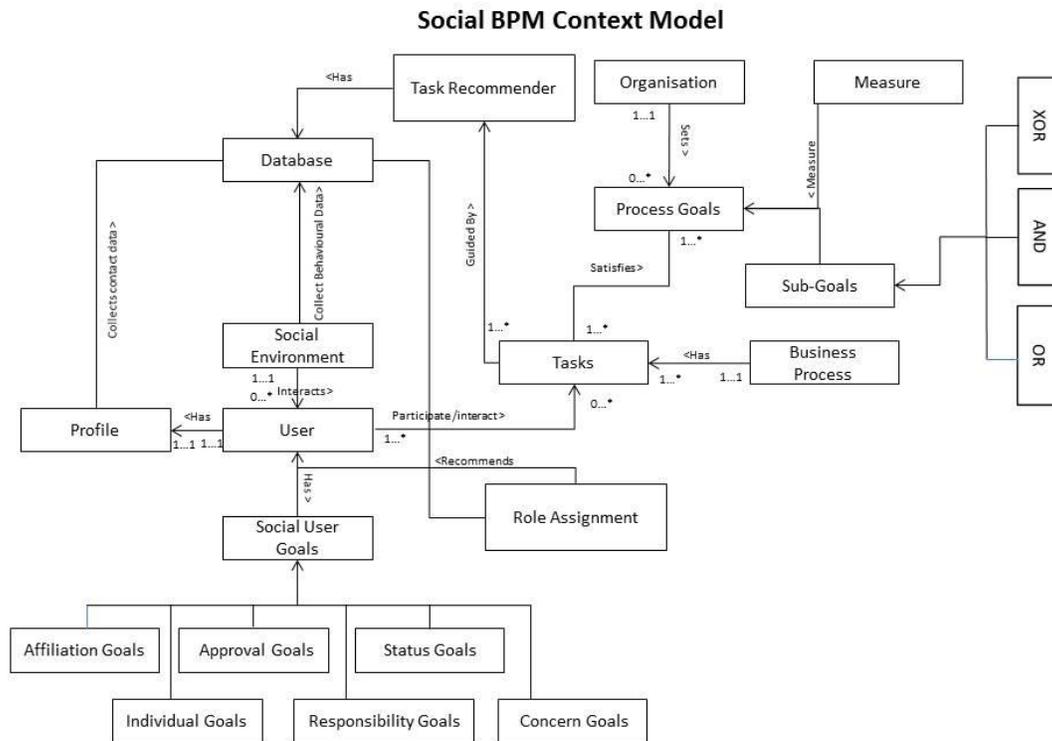
Fig.1. High-level model of the proposed social BPM framework

The interactions take place in the light of the social environment, which the process recommendation mechanism and role assignment function depend on.

The recommendation engine proposes the various goals which need to be fulfilled to the user and makes these recommendations based on the course of action the user takes.

The role assignment functionality recommends the most suitable user to be responsible for the overall execution of the processes to achieve the specific goals. The method of rating the suitability of the user is explained in later sections.

A typical recommender system has 3 main elements to it, 1. Background data which is the information the system has stored about the user (This is equivalent to the behavioral, profile and feedback data in our proposed model [6], 2. Input data which is the data given to the system as items to be recommended/assigned (Similar to the projects available which need to be recommended to the most suitable user), 3. Finally an algorithm that combines the background and input data to produce an outcome which is the assignment/proposal of the two data together [21].

IV. TASK RECOMMENDATION IN SBPM

Recommender systems became popular and an important area of research since the emerging of researches in collaborative filtering in the mid-90s [86]. Recommender systems [20] [19] have been mainly used in e-commerce, content presentation, entertainment and services [19], however they have not been used previously in the context of SBPM.

The recommendation which is discussed here is in the context of social processes, and used as an approach in order to benefit from the process knowledge captured in previous similar business process executions, so it does not use the same algorithms used in *Content-based*, *Collaborative Recommendation and Hybrid approaches* [18].

A. Tagging

Tagging involves the assignment of unrestricted keywords to all kinds of content [11]. These assignments can assist to build up a knowledge repository which contains related useful information on a certain topic. Tagging, for the first time has been proposed to be used in the context of business processes in [10]. This is in order to capture any information used in process execution, and to make these available in the future. This is done through a tagging mechanism of the discussions which take place for the fulfillment of the tasks in a specific process. Social Business Process Management can therefore benefit from such approach to assist and enable improvement to the processes.

Researches [13] in tagging have also categorized tags into different types such as context-based, content-based, subjective and organizational which essentially define what the tags refer to it.

B. Social Tagging Archetecture

Expanding on previous research [10] carried out in this area, here we define three types of tags which are captured by the system. After the completion of the tasks in any given process, the discussion can be tagged with one or more of the following types:

<<Process>> : this tag captures the general process in which the discussion is taking place.

<<Task>> : this captures the task which the discussion is about in order to fulfil the overall process goal.

<<Rationale>>: this tag captures the reason and rationale behind tagging a particular segment of the discussion.
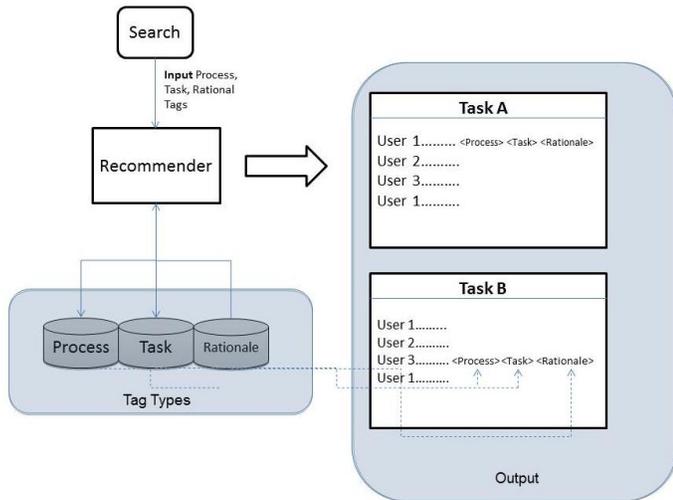
Fig.2. Social Tagging in SBPM

This tagging is done by a process owner [7] who is recommended and chosen based on a number of criteria (interest, expertise and feedback) and is responsible for tagging the discussion with these tags. As demonstrated in fig 2, there is discussion which takes place by the social user community as part of fulfilling each of the tasks. The process owner who is overall responsible for the completion of the process then tags the discussion which can be useful in future executions. After the completion of the tagging which can be quite challenging depending on the domain, the user who would like to initiate or take part in a similar process in the future is able to utilized the previously captured knowledge by searching a number of key words.

Fig 3 illustrates how the search can be one or more than the type of the tags mentioned above. The user might only be interested in retrieving information which is relevant to the rationale he would like to follow, or he can be more specific and customize the process domain he wishes to look for in which case he will mention the type of the 'process' and 'tasks' which he would like to search.

Based on the input, the system searches through the tags and returns the discussions related to the terms. The discussion and key points mentioned can then be used in the current process which the user is about to run.

With the increase of the number of execution logs and the tags in the system, the process knowledge is expanded and becomes more rich. This will assist the users in the future who would like to partake in similar type of processes.
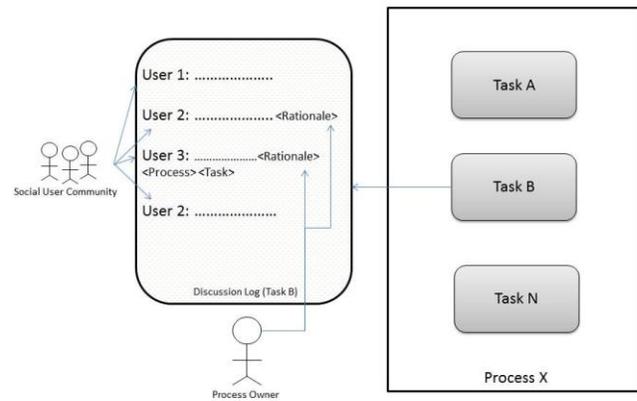
Fig.3. Process Knowledge Discovery in SBPM

During the discussion, there is a real-time interaction and catering for unplanned participation from the user community which is essential for the SBPM model [9], this interaction model is presented in [6]. The purpose of this is to allow flexibility and a degree of freedom to the user in carrying out the tasks and the set goals.

C. Proposed Meta-Model for Social Tagging in BPM

Having explained the mechanism of the task recommender and how it uses social tagging to enable process knowledge discovery, fig 4 presents a meta-model of the proposed approach towards social tagging in the context of social BPM. Essentially, the process owner is a user which is responsible for the tagging of the discussions. These are tagged with different tag types as explained above, and stored in a tag cloud. The user can then use the tag could and enter free text to indicate the type of process of area he is interested in and retrieve the discussion related to the searched terms.
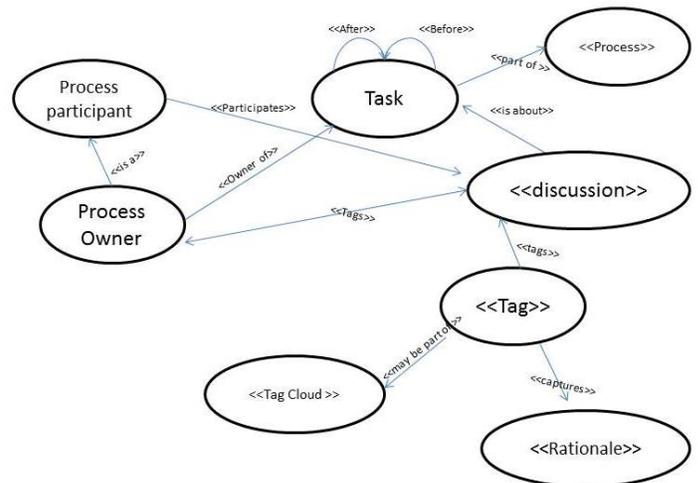
Fig.4. Proposed Meta-Model for Social Tagging

The traditional BPM models do not support the required flexibility and agility for SBPM when it comes to rapid changing and unpredictable process steps; this makes flexibility of the processes during runtime impossible [15].

Therefore our proposed recommender process aims to provide a level of flexibility based on the users course of action and guide the user until all the goals are fulfilled. In other words, the users benefit from previously captured process knowledge in order to fulfil the required goal of the process in a semi-structured way without having the process steps imposed on them.

D. A Worked Example

To illustrate the theory explained above, consider the process of organizing a football tournament. This process consists of tasks such as booking a pitch, deciding on a time and venue, finding a referee and organizing the match fixtures and so on. After all of these tasks have been collectively accomplished, the process owner revisits the discussions related to each one of these tasks and tags relevant fragments which could be utilized in the future when organizing similar events.

For instance, for the pitch site and venue, after investigating a number of locations and different pitches, the organizing team decided to go with astro turf option. In such an instance, the process owner can tag that segment of the discussion about the choice of the venue with a tag such as 'AstroTurf'. This would be the rationale behind the tagging, he can also use other tags such as 'pitch' to indicate the task which he is tagging and also add the tag 'FootballTournoment' to specify the type of process this tag is for.

Fig 5 demonstrates what has been explained by showing how the tagging will take place after the completion of each of the tasks. The same process will follow for each of the other tasks as well, the rationale maybe different in each case and is left to the process owner to decide. The following is only an example for finding a pitch.
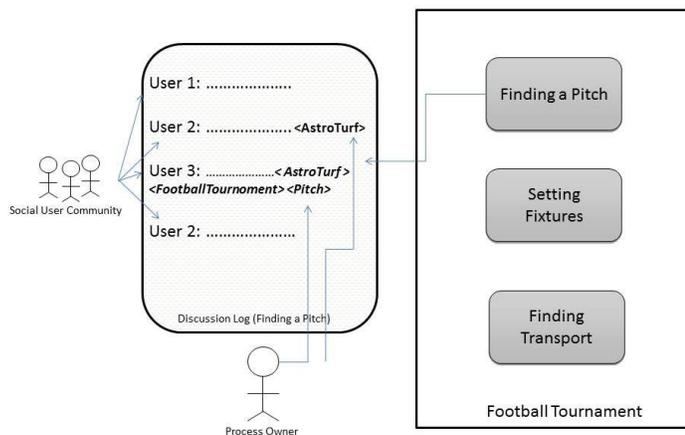
Fig.5. Example of social tagging for a football tournament process

Furthermore, the discussion for the pitch selection can be linked to the discussion and task related to the transportation method to get to the venue. So in the future when someone searches for that specific type of venue, after finding the exact pitch which was decided on, the user can also see the easiest/cheapest/quickest (depending on the tag used for the transportation) to the venue and consider using the same way without the need of having to do the research from scratch.

Thus, not only domain specific knowledge ('AstroTurf) is retrieved but also process specific knowledge (i.e. how to travel to that particular location). The more instances of a process are executed and annotated, the higher the amount of knowledge that will be available for future users to utilize. For their more efficient utilization, previously used tags can be stored in a 'tag Cloud' [17] where the users can see what tags have been previously and their frequencies of use.

V.   CONCLUSION AND REFLECTIONS

Social BPM research has developed over the past few years and preliminary meta-model [30] has been proposed. This paper unfolded the main components of SBPM which is the task recommendation mechanism to support the user during process execution.

Goal-based approach has been adapted as a way forward for social BPM, the reason for this simply due to the flexibility and control that this approach can bring to the business process management lifecycle.

Goal-oriented SBPM framework is informed by the role-assignment engine and the task recommendation mechanism. The latter was explained in detail throughout this paper, demonstrating how the user is able to benefit and utilize the knowledge captured in past executions. This is done through introducing social tagging to the SBPM model to enable the capturing of process knowledge. These can then be used by users when searching based on a specific criteria they are interested in.

Such an approach truly shifts the BPM model to a socially driven one which is guided by its social context of the users and tasks performed in the past. There is no imposing in such a system and the user and process owners are flexible in the course of actions they wish to take in order to fulfil a specific task or benefit from the captured process knowledge. It was argued that by utilizing the power of social tagging after the completion of the process, future users can benefit from wealth of process related ideas and approaches and reuse them as appropriate. The tagging of the tasks and discussions about them, adds to the whole process knowledge.

The proposed model has also its own limitation which will be addressed in future research. Some of these challenges start from the initial stages of setting the tasks and in the case of conflating goals [25] with tasks (which will reduce the flexibility given to the user), or defining conflicting goals are examples of problems which this approach could face.

The other limitation of the use of social tagging in this context is, the tedious tasks of the process owner tagging the discussion and going through what has been discussed line by line to identify a rationale for some of the segments and tag them. Although the main use of having the <process> and <task> tags are primarily to enable customization during searching, it would be straightforward to do in the text. However identifying the rationale could be challenging and time consuming.

Different approaches such as 'Fuzzy Logic' [14] could potentially be adapted to be applied in order to assist in the tagging process. Although the accuracy of the tagging might not be the same as manual tagging, however it is an area where further research can be carried out to investigate whether any approach can be used to assist this process.

Furthermore, the problem with free text searching and mismatches has always existed in search engines. This is no different, so although the search mechanism will look through the tagged terms, and find the closest and most relevant match. However, if the incorrect term is used, the user might not be able to get the result her has been looking for. We will try to investigate this further and overcome some aspects of this issue by storing the tags in a tag cloud and make available to the users, so they are able to see all the tagged terms and used them in order to retrieve the process knowledge they are looking for.

Overall, the proposed approach is a step closer to a comprehensive social BPM model which researchers have been talking about for the past few years. Social tagging is used extensively in social media and different social platforms, but it has not been adapted in the context of business process. This model brings about great benefit to the process users and designers and to the overall design, implementation, enactment and evaluation of the business process lifecycle.

The proposed approach towards process knowledge discovery and the role-assignment [7] mechanism need to be validated and applied practically to a real-case study. For this purpose we are in the process of producing a prototype using a wiki in order to implement what has been proposed.

REFERENCES

[1] Kemsley, S.: Leveraging social BPM For Enterprise Transformation: In Social BPM Work, Planning and Social Collaboration Under the Impact of Social Technology, Swenson, KD., Palmer, N., et al.: BPM and Workflow Handbook Series, pp. 77 – 83, Future Strategies Inc, USA (2011).

[2] R.G. Lee, B.G. Dale, "Business process management: a review and evaluation", Business Process Management Journal, Vol. 4 Iss: 3, pp.214 – 225, MCB UP Ltd (1998).

[3] Brambilla, m., Fraternali, p., Combining social web and BPM for improving enterprise performances: the BPM4People approach to social BPM, In Proceedings of the 21st international conference companion on World Wide Web, pp. 223-226 ACM Publishers, New York (2012).

[4] Filipowska, A., Kaczmarek, M., Koschmider, A., Stein, S., Wecel, K., Abramowicz, w., Social Software and Semantics for Business Process Management - Alternative or Synergy?, in Journal of Systems Integration, vol 2, No 3, pp.54 - 69 (2011).

[5] S. Erol, M. Granitzer, S. Happ, S. Jantunen, B. Jennings, A. Koschmider, S. Nurcan, D. Rossi, R. Schmidt, P. Johannesson. Combining BPM and Social Software: Contradiction or Chance? In Journal of Software Maintenance and Evolution: Research and Practice, vol. 22 pp. 449–476, John Wiley & Sons, Ltd (2010.)

[6] Rangiha, ME., Karakostas, B.: A Goal-Oriented Social BPM Framework. (Submitted) In: Post-Proceedings of Lecture Notes in Business Information Processing (LNBIP) series, Springer, 15th BPMDS2 workshop in conjunction with CAiSE 14 conference, Thessaloniki, Greece, China (2014).

[7] Rangiha, ME., Karakostas, B.: Process Recommendation and Role Assignment in Social Business Process Management. In: Proceeding The Science and Information Organization, Science and Information

(SAI) Conference, 27 - 29 Aug (2014).

[8] Swenson, K.: Who is Socializing in Social BPM? 12 May 2010 [Blog Entry], Collaborative Planning & Social Business. Available: http://social-biz.org/2010/05/12/who-is-socializing-in-social-bpm-2/ [Accessed Sep 2014].

[9] Seybod, P.: The next big thing: Adaptive Business Process Management, Users.com, [Blog Entry]: Available at: http://www.users.com/articles/the-next-big-thing-adaptive-business-process-management/ [Accessed Oct 2014].

[10] Rangiha, ME., Karakostas, B.: The Use of Social Tagging in Social Business Process Management.(In Print) In: Conference Proceedings The International Conference Business Process Management, London, UK, 12 - 14 Oct (2014)

[11] Cantador I., Konstas I., Jose J.: Categorising Social Tags to Improve Folksonomy-based Recommendations.: In Journal Web Semantics: Science, Services and Agents on the World Wide Web, V.9, Iss. 1, pp. 1 – 15, March (2011).

[12] Rich, E.: User Modeling via Stereotypes. In: Cognitive Science 3, pp. 329–354 (1979).

[13] Panke, S., & Gaiser, B.: With My Head Up in the Clouds: Using Social Tagging to Organize Knowledge. Journal of Business and Technical Communication, 23(3), 318. (2009)

[14] Lotfi A.Zadeh: Is there a need for fuzzy logic ?.: Science Direct: In information sciences 178, Elsevier Inc, pp. 2751 - 2778 (2008).

[15] Burkhart, T., Loos, P.: Flexible Business Processes - Evaluation of Current Approaches: MKWI '10, Institute for Information Systems (IWi) at the German Research Center for Artificial Intelligence (DFKI. Göttingen (2010).

[16] Richardson, C., Is social BPM a Methodology, A Technology, Or just a lot of Hype? 20 May 2010 [Blog Entry]. Forrester. Available: http://blogs.forrester.com/clay_richardson/10-05-20-social_bpm_methodology_technology_or_just_lot_hype[Accessed Oct 2014].

[17] Surowiecki, J. The Wisdom of Crowds. Abacus; New Ed edition, 1st edition, (2005).

[18] Adomavicius, G.: Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. In: IEEE Transactions on knowledge and data engineering, vol 17, no 6 (2005).

[19] Koschmider A., Oberweis A.: Designing Business Processes with a Recommendation-Based Editor, Handbook on Business Process Management 1, pp.299-312 (2010).

[20] Balabanovic, M., Shoham, Y.: Fab: Content-based, collaborative recommendation. In: Communications of the ACM, 40(3):66-72, USA, New York (1997).

[21] Burke, R: Hybrid Recommender Systems: Survey and Experiments: In Journal User Modeling and User-Adapted Interaction, vol 12, issue 4, pp 331-370, Kluwer Academic Publishers Hingham, MA, USA (2002).

[22] Kavakli, E.: Goal-Driven Requirements Engineering: Modelling and Guidance. Available: http://www.ct.aegean.gr/people/vkavakli/publications/pdf_files/PhD/kavakli_chapter2.pdfhype [Accessed Sep 2014]

[23] Yu, E.S.K. and Mylopoulos, J.: Why Goal-Oriented Requirements Engineering. Proceedings of the 4th International Workshop on Requirements Engineering. Foundations of Software Quality (REFSQ '98), pp. 15 – 22. Italy, Pisa (1998).

[24] Dellen, B., Maurer, F., Pews, G.: Knowledge based techniques to increase the flexibility of workflow management. In: Data and Knowledge Engineering, North-Holland (1997).

[25] Adams, M., Hofstede, A., Edmond, D., van der Aalst, W.: Facilitating Flexibility and Dynamic Exception Handling in Workflows Through Worklets. In: Proceedings of the CAiSE '05 Forum, pp. 45–50, FEUP (2005).

[26] Rangiha, ME., Karakostas, B.:Goal-Driven Social Business Process Management. In: Proceeding The Science and Information Organization, Science and Information (SAI) Conference, pp. 894 – 901, London, UK (2013).

[27] Nurcan, S., Etien, A., Kaab, R., & Zouka, I.: A strategy driven business process modelling approach. Journal of Business Process Management, 11, 6, pp. 628-649 (2005).

[28] Filipowska, A., Kaczmarek, M, Kowalkiewicz, M., Zhou, X., Born, M.: Procedure and guidelines for evaluation of BPM methodologies: In Business Process Management Journal, vol. 15 Iss: 3, pp.336 – 357 (2009).

[29] Dellen, B., Maurer, F., Pews, G.: Knowledge based techniques to increase the flexibility of workflow management. In: Data and Knowledge Engineering, vol 23, Issue 3, pp. 269–295, North-Holland (1997).

[30] Rangiha, ME., Karakostas, B.: Towards a Meta-Model for Goal-Based Social BPM. (In press) In: Post-Proceedings of Lecture Notes in Business Information Processing (LNBIP) series, Springer, 6th BPMS2 workshop in conjunction with BPM 2013 conference, Beijing, China (2013).

[31] Kueng, p., Kawalek, P.: Goal-Based Business Process Models: Creation and Evaluation: In Business Process Management Journal Emerald, vol.3, no.1, pp. 17-38. MCB University Press (1997).

A new algorithm for detecting SQL injection attack in Web application

Ouarda Lounis, Salah Eddine Bouhouita Guermeche, Lalia Saoudi, Salah Eddine Benaicha
Computer Science Department
University of Mohamed Boudiaf of M'Sila
M'Sila, Algeria

Abstract—**Nowadays, the security of applications and Web servers is a new trend that finds its need on the Web. The number of vulnerabilities identified in this type of applications is constantly increasing especially SQL injection attack. It is therefore necessary to regularly audit Web applications to verify the presence of exploitable vulnerabilities. Web vulnerability scanner WASAPY is one of the audit tool, it uses an algorithm which bases on a classification techniques of pages obtained by sending HTTP requests especially formatted. We propose in this paper a new algorithm which was built in a vision to improve rather to supplement the logic followed in modeling WASAPY tool. The tool was supplemented by a new class reflecting the legitimate appearance or referential, therefore, the detection mechanism was solidly built on a statistic in a fairly clear mathematical framework described by a simple geometric representation or interpretation.**

Keywords—*SQL injection attack; scanner Web; Web Application; Web vulnerabilities; security*

I. INTRODUCTION

The Web server security is now a recurring problem. The number of vulnerabilities identified in this type of software is constantly increasing, as described in particular in the document "The OWASP Ten Most Critical Web Application Security Risks" [4]. They can be explained by several reasons: the increasing complexity of Web technologies, deadlines ever tennis marketing software, sometimes limited skills and lack of safety culture developers. As a result, many of these applications contain multiple vulnerabilities that can be exploited by hackers. These attacks can allow them, for example, to obtain confidential data (credit card numbers, passwords, etc..) that are manipulated by the application, or even alter or destroy some of these data. The complexity of the technologies used today (Java, JavaScript, PHP, Ruby, J2E, etc..) to create Web applications makes it particularly difficult 1) to prevent the introduction of vulnerabilities in these applications and 2) estimate or predict their presence. In addition, network security and installation firewall does not provide adequate protection against Web attacks as these applications are accessible to all. It is therefore necessary to regularly audit Web applications to verify the presence of exploitable vulnerabilities and this can be realized by vulnerabilities Web scanners.

There are two main classes of approaches adopted by most of the vulnerability Web scanners:

- Approach based on recognition of error messages in response pages,

- Approach based on studying similarity of pages returned by the server.

These two approaches will be explained in section III.

In our project, we realize an approach for the detection of SQL injection attacks in Web applications, based on sending HTTP requests and analyzing the responses of the latter. This approach is based on two techniques: technique of recognition of error messages in response pages, and the study of similarity of pages returned by the server.

The proposed algorithm is a modified and improved approach of Rim's Akrout [6] algorithm, it is carried out in a spirit of logical completeness expressing a desire to specify a new model completing the WASAPY model. It involves sending HTTP requests to the Web server. These requests are structured into three classes: class containing syntactically valid requests, class containing syntactically invalid requests and class containing random requests. The responses of each request will be recorded and compared to a fourth class that contains only pages of references to examine the similarity between them using the Levenshtein algorithm [5], and following tests, we determine what are the requests that have well used to exploit the SQL injection vulnerability.

The LS model provide a detection mechanism built on a solid statistics in a clear mathematical framework covering the whole crowd resulting data. Detection of security following two phases 1) detect if the page is secured or not, and 2) if the page is not secure, a search request injection is performed.

Our scanner is experienced using vulnerable Web sites we implemented to calculate the performance of the approach implemented, according to three parameters: the detection rate, false positive, false negative. Graphs are drawn to properly express the results of experiments realized. The used request were extracted from OWASP (SQL Cheat Sheet) project [4].

II. RELATED WORK

A. using the recognition of error messages in response pages approache

W3af [1] was created by Andres Riancho in 2006 and is considered one of the most powerful scanners, it is written in Python. Its modular architecture allows users to import and easily modify the different modules that compose it. W3af sends three HTTP requests to test the presence of a vulnerability in a page, the three responses associated with these queries are then analyzed. If they contain SQL error messages, w3af informs the user that the application is vulnerable to SQL injection. In spite of its power, it has no additional mechanism that is implemented to verify if the vulnerability actually exists or not, that is to say, if it is actually exploitable, this is its major default.

Wapiti [2] is another example which follows the same principle. This tool developed in Python, is able to detect SQL injection, XSS injections, mishandling of files, LDAP injection and execution of operating system commands from a URL. To identify SQL injections, it sends two requests. Vulnerability is declared present if an error is identified in the response pages. The effectiveness of this approach is related to the completeness of the knowledge base regrouping the error messages.

B. using the analysis similarity of pages returned by the server approache

Skipfish [3] was developed by Google to detect vulnerabilities on Web servers. It proceeds in two steps. In a first step, he analyses the Web application and collect all pages that appear to be stable. The others are ignored. To detect whether a page is stable Skipfish sends 15 queries and several tests are performed on the page. The main shortcoming of this scanner is that the distance used for the study of similarity considers the frequency of words regardless of the order of words in a text. Ignore the word order can lead to ignore the semantics of a page and again can lead to misjudge if two pages are identical or not. For example, the following pages share the same vocabulary, but they correspond to a successful and failed authentication respectively:

Your are authenticated, you have-nots has Entered wrong login.

Your are not authenticated, you-have Entered a wrong login.

III. EXPLANATION AND CRITICISM OF APPROACHES ADOPTED BY WEB SCANNER

A. Approache based on recognition of error messages in response pages

To identify SQL injections, this approach sends requests of a particular format and look for specific patterns in responses such as error messages database. The basic idea is that the presence of an SQL error message in an HTML page response means that the corresponding request has not been verified by the Web application before being sent to the server database. Therefore, the fact that the request was sent unchanged to the

SQL server reveals the presence of a vulnerability. Scanners such as w3af (SQLI module) andWapiti adopt such an approach.

The effectiveness of the approach by recognition of error messages is related to the completeness of the knowledge base regrouping the error messages that may result from the execution of queries submitted to the Web application. Generally, as in the case of w3af we consider mainly the error messages from the database. However, the error messages that are included in the HTML response pages do not necessarily come from the server database. The error message may also be generated by the application that can also reformulate the error message from the server, for example to make it understandable to the client. Moreover, even if the message is generated by the database server, the receipt of this message is not sufficient to say that SQL injection is possible. Indeed, this message means that for this particular query, entries have not been cleaned, but this does not support the conclusion in relation to other SQL requests, particularly those correspond to successful attacks.

B. Approache based on studying similarity of pages returned by the server.

The principle of this approach is to send various requests specific to the type of vulnerabilities and to study the similarity of the responses returned by the application using a textual distance. Based on the results obtained and well-defined criteria, we conclude on the existence or not of a vulnerability.

Concerning this approach, it is based on the assumption that the contents of rejection page is generally different from the contents of execution page. For this comparison, however, to be effective, it is important to ensure a wide coverage of different types of pages of rejection that could be generated by the application. This can be achieved by generating a large number of queries to enable the largest possible number of different pages of rejection. However, existing implementations of this approach, especially in Skipfish generate too few queries. Skipfish only uses 3 queries. If the answers correspond to different pages of rejection, he incorrectly concludes that the vulnerability is present leading to a false positive. Moreover, this approach, as in any classification problem, the choice of the distance is very important. That used in Skipfish do not take into account the order of words in a text. However, this order generally defines the semantics of the page. It is therefore important to consider the order of words in text.

These analyzes clearly show the need to develop new approaches to improve the effectiveness of vulnerability tools detection. The work presented in this paper adopt a new approach.

IV. GLOBAL PREVIEW OF OUR APPROCH

Security is a concept which requires a certain reflection especially if we want to give it a definition which allows us to develop methods or techniques allowing to handle certain quantities which capture as faithfully as possible the intrinsic aspects of the security of the system in question.

Our study is interested in the security of the Web application, of which it is necessary to specify the necessary measures which can reflect the state of its safety.

To do this, it seems to us essential to have a model vision of the security of Web application generally.

To talk about safety is to call upon the legitimacy and thus security and legitimacy are two equivalent notion in the following optic:

To be secured is to allow only the legitimate attitude which can in this case of Web application, be seen as a natural navigation in the Web application such as designed. So, a navigation is defined as a sequence of legitimate Web pages which can be grouped within a single class called class Ref.

And therefore the legitimacy can be defined as a passage of a page of reference to another reference page, in this vision of things, a secure Web application generate only legitimate pages and thus it belongs in the reference class. All in all, there is a unique class being invoked. And thus legitimacy or security is modelled by the unity described by a unique class containing all the reference pages of the Web application.

Conversely, the illegitimate has a strong ability to generate non-legitimate pages, evidently they don't belong to the class of reference pages Ref, and so it proposes a new classes. This fact, the illegitimate which captures the state of a not secure Web application; is described by an explosion of the unique class in new classes.

These new classes are created according to a criterion merging three criteria simultaneously. These criteria stemming from the syntactic correction and the sematic of the SQL.

- Syntactic aspect is controlled by the SGBD which generates errors messages suited to the produced syntactic violation.

- Values of legitimate attributes generate pages slightly different from references, while those with illigimate values produce error messages generated by the SGBD.

- The semantic aspect is present with the insertion at the request of tautologies or antilogies.

The fission which reflects the non-secure of the Web application considered produces two new classes:

- Class Aleat: contain the pages returned by the random requests.

Example : Considering the URL following with the parameter: id_suj

http://localhost/forum/ajouterreponse.php?id_suj=254

here, the id-suj value is randomly chosen.

- Inval class: contains the pages returned by the invalid requests.

Example:

http://localhost/forum/ajouterreponse.php?id_suj=1union privilege_type FROM information_schema.user_privileges WHERE privilege_type = 'SUPER' AND 1=1

The definition of classes is based on two fundamental concepts:

- The inside-class coupling: the elements of single class should be very close one of the other one, what reflects the intuition "those who are alike flock".

- The between-classes decoupling: the elements of two different classes should be remote one of the other one, what reflects the intuition "the remoteness disadvantage the union".

The coupling and the decoupling are defined on the notion of distance. In fact, this distance capture the degree of similarity between elements. The similarity is defined mathematical as a function acting on two HTML pages and returns a real number in the interval $[0,1]$. Thus, the mathematical sign of the similarity is as follows:

$$\textbf{Sim: } C_1 \text{ x } C_2 \qquad [0,1]$$
$$\textbf{Sim: } [P_1, P_2] \qquad s$$
$$\textbf{Sim } (P_1, P_2) = s$$
$$\textbf{Sim: Similarity}$$
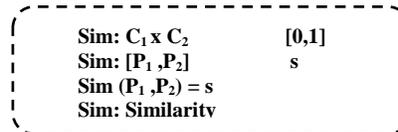
Fig. 1. Similarity between two pages

Many researches were deployed for the definition of functions of similarity between two strings. Their computation strategies differ according to their vision of the string:

- *Character by character:* The matching involves both a character of the first with the character of the same order but took the second page. Among the most famous is the Levenshtein distance

- *Line by Line:* The matching involves two lines each time, each taken from one of the pages, one of the most famous is the diff Linux.

- *Word by Word:* The matching is done by comparing a word of the first page with the corresponding word in order in the second page. Taking into account the order of the words in each page define two variants of this strategy:

  - With consideration of the scheduling of words.

  - Without consideration of their mutual scheduling.

In our case, we used Lenvenshtein algorithm [5] which will be described in the fifth section.

The vulnerability if it exists, it sounds from two points called injection points:

- URL: is identified by the presence of URL parameter.

- Formulary: is identified by text fields such as the login and the password.

V.   THE WASAPY MODEL

A model recently proposed within the framework of a Phd thesis in the national institute of applied sciences of Toulouse (INSA Toulouse) [6].

A. *Presentation of the model*

a) *The vision of safety:*

The security of Web application is seen as the non-presence of illegitimate pages generated by a malicious behavior. Thus, the model articulates on the illegitimate to explore the security status of the Web application in question. This way of making influences considerably the modelling of the approach.

b) *The classes of the model:*

The model proposes two classes: the class Aleat of Web pages generated by random requests as well as the class Inval including Web pages turned by invalid requests. We notice well that the choice of the classes reflects the spirit of the approach since the two classes represent the malicious requests. In addition to these two classes, we found a class which presents a syntactically valid requests.

c) *The principal of detection:*

The detection principle is formed around the clustering technique implies a similarity threshold ε defined as follows:

- The ε algorithm

Set the threshold by the smallest distance between i) The longest distance between two responses in Aleat,

ii) The longest distance between two in the invalid responses.

- The clustering algorithm

Grouped together within the same cluster those queries which the pair wise distance is less than a threshold. [7]

- Detection Algorithm

If Ǝ a Grap whose members are only valid requests Then all these request have allowed to exploit the injection, so the site is not secure.

Fig. 2. Detection algorithm

B. *Criticism of WASAPY model*

a) *The vision of security:*

The vision was bounded on the illegitimate part of the queries, which deprived the completeness of a model that could be very beneficial. This gap has greatly influenced the accuracy of the proposed approach. Also, the model classes were limited to two classes reflecting malice only at the expense of the correction.

b) *The classes of the model:*

The class Aleat and the class Inval should respect the classification ethics which are:

- A coupling factor: as low as possible which expresses assembling similarity. This factor is well respected by tests conducted in the study of WASAPY model.

- A factor of Decoupling: most important factor that justifies the creation of new classes. However, an interesting observation draws our attention to the fact that the decoupling factor is so low that coincides with the coupling factor. This reflects negatively on the quality of classes offered, because of the similarity point of view, the two classes are only one.

c) *The choice of threshold ε:*

The choice of ε was motivated by the desire to minimize false positives products, and therefore, its design foundation was not mathematically or statistically very clear. The evaluation of the threshold ε was much guided by intuition that needed empirical experiments to acquire a certain degree of confidence.

d) *Detection:*

Detection is significantly related to the threshold ε, the design of the latter will have an impact, especially, the fact that it was designed with the intent to minimize false positives.

VI. LS MODEL

A. *The vision of security*

The vision of security is induced by the completeness of legitimate and illegitimate space. And thus, to be secure is to be closer to legitimate than illegitimate.

B. *The model classes*

a) *The classe Ref*

Reflects the legitimate website navigation, which is ultimately a set of legitimate pages called "reference pages" that are offered by the site during normal navigation. And so, a page that is not offered by a natural navigation of the website is recognized as a page raising from a malicious attempt to access unauthorized information and so this page is an attack signature that can reveal a vulnerability in the web site design. These pages are the result of some queries which represent a navigation parallel to that offered by the site navigation.

b) *The class Aleat*

Aleat is the class which queries are generated from words randomly selected from the list [a- zA -Z0 -9]. These queries generate pages rejection [6].

c) *The class Inval*

Inval is the class of SQL injection, syntactically invalid queries to the injection point. They are constructed so that the SQL Server that interprets these queries generates an error [6].

d) *The class Val*

Val is the class of queries that generates execution pages. For example, the couple login / pass which respectively have the values ' or'1 ' = '1 - abcd and generates the following SQL query :

SELECT * FROM users WHERE login=' 'or'1'='1--' and pass='abcd'

C. *The principle of detection*

The detection principle is formulated through two phases :

- Determine if the site is secure or not.

- If the site is not secure , find the maximum injection queries.

a) Phase one :

A secure page allows only legitimate consultation therefore it returned legitimate pages or their ones which are very closer. The neighborhood is defined with the distance which is a similarity function. This is a measure that quantifies the degree of similarity between two pages. So securing entails approximating pages returned by different queries from the three classes of LS.

However, if the page is not secure, then it has vulnerabilities that can be exploited by illegitimate requests, applied to injection points producing illegitimate pages that are significantly distant from legitimate pages. And therefore the none-securing induce removal of pages returned. The speed away of, depend from the classes of the queries.

The notion of proximity is quite relative, which should be quantified in order to allow a specific numerical language much more than descriptive one. The nature of the analysis of such studies provide a huge amount of data , and take into account all the information that its confine is a challenge that can be overcome with a good statistic . The latter should be defined in a way that to get the most information in the most compact possible form.

Descriptive statistics gives us a simple tool which is the arithmetic mean. It can be very effective if properly exploited, especially if it is engaged in a statistics of levels.

The model LS formalizes detection with a triangular vision focuses on the three classes in a space of four classes: Ref, Aleat , Inval and Val . The detection formula is described with highly descriptive geometric language. The following diagram describes geometrically the first phase of detection principle:
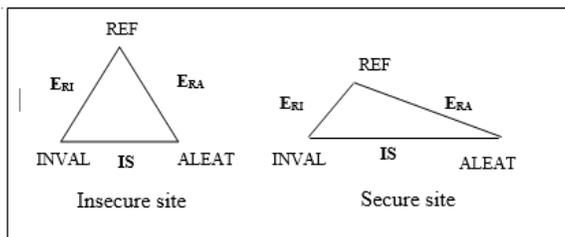
Fig. 3.   Triangular representation of secure and non-secure page

E_{RA}: expectation between the similarity between the reference pages and answers random queries.

E_{RI}: expectation between the similarity between the reference pages and answers invalid queries.

IS: Security Index.

The principle of securing is as follow:

If (triangle's base < ½ (sum (ERA , ERI)))
Then the page is not secure
Else it's secure.

Fig. 4.   Principle of securing

The detection principle of securing a page is formulated on a strong enough bases borrowed from descriptive statistics. The measure (expectation) is simple but effective to capture the essential and relevant information for our study.

a) Phase Two: injection queries?

In the case where the detection phase determines that the page is not secure, it engages a set of queries to find those that can exploit SQL injection. The principle of detection of injection queries is formalized following the same spirit of reflection already expressed in the first phase detection.

Illegitimate applications of both classes: Aleat and Inval return pages that are clearly distant from legitimate pages of the class Ref. Knowing that the pages of the two classes : Aleat and Inval who share the same offense of illegitimacy making less distant from each other compared with the class Ref. We conclude that a non-secure page is sensitive in its reaction with illegitimate requests without paying much attention to their syntactic correctness.

Similarly, here too, the concept of proximity is formalized with a borrowed geometry in a triangular vision language represented as follows:
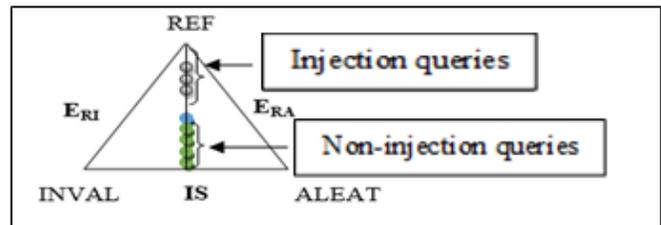
Fig. 5.   Triangular representation of the vulnerability detection on non-secure page

D. The singularity of LS Model

The singular point of the model is: IS = 0 which means that the basis of reasoning [Inval, Aleat] disappeared. Therefore, it does not revolve on what to make a decision to secure, because the responses of invalid and random queries are merged into a single class (decoupling factor = 0, coupling factor = 0). From the triangular vision, we go at the linear vision, we don't need references, and the two classes Aleat and Inval are no longer discernible. The role of class Ref became neutral.

And therefore, the need for valid class arises. The geometric description of the model became a line whose ends are fused two points (Aleat , Inval) and the second is (Val) . If it retains its geometric shape of this line expresses the non-secure page and if it shrinks to a single point is to express the security and confirm our starting view that unity expresses security and bursting none securing. This singularity of the model represents the SQL injection applied to authentication forms (POST).

E. Algorithm global of detection

S: is the distance between two textual responses (Levenshtein algorithm).

ReqAleat, ReqInval, Ref, ReqVal: are respectively the random queries, invalid, reference pages and valid requests.

RepAleat, RepInval, RepRef, RepVal: are respectively random, invalid, reference pages and valid responses.

```
    Integer i:= 0;
    ReqAleat, ReqInval, Ref, ReqVal;
    Read responss of each type of the queries: RepAleat,
RepInval, RepRef, RepVal;
    Vector B:= Dis (RepAleat, RepInval); Vector A1:=
Dis (RepRef, RepInval;
    Vector A2:= Dis (RepRef, RepAleat); Vector C1:=
Dis (RepVal, RepAleat);
    Vector C2:= Dis (RepVal, RepInval); Vector D:= Dis
(RepVal, RepRef);
    Calculus : IS = avg(B); Calculer : x:= avg(A1,A2);
    If (x < IS) the page is secure End If
    If (x > IS)
        Loop
          Calculate c:= avg(C1i,C2i);
          If (c > IS) the ith query of the valid queries
         has    successes to exploit un SQL injection.
          Else the query has not permit the exploitation
         of   the injection.
          i:= i+1;
        Until  i= NbrReqVal
    End If
    If (IS = 0)
        If (C1 =0 AND C2 =0)   the page is secure
        Else the page is not secure.
    End If
```

Fig. 6.   Global algorithm of detection of the LS model

```
    integer DistanceOfLevenshtein(string chaine1, string
chaine2)

    // i and j iterate over  chaine1 et chaine2
    Integer  i, j, cost, lengthChaine1, lengthChaine2
    lengthChaine1 := lengthr(Chaine1)
    lengthChaine2 := length(Chaine2)
    // d is a table of  l engthChaine1 +1 rowws and
lengthChaine2 +1 columns
    Integer d[0..lengthChaine1, 0..lengthChaine2]

    for i from 0 to lengthChaine1
       d[i, 0] := i
    for j from 0 to lengthrChaine2
       d[0, j] := j

    for i from 1to lengthChaine1
       for j from 1to lengthChaine2
         if chaine1[i - 1] = chaine2[j - 1] then cost := 0
         else cost := 1
         d[i, j] := minimum(
                      d[i-1, j  ] + 1,    // supression
                      d[i,  j-1] + 1,    // insertion
                      d[i-1, j-1] + coût  // substitution
                    )

    return d[lengthrChaine1, lengthChaine2]
```

Fig. 7.   Levenshtein algorithm

The first two cases (x> IS) and (x <IS) formalize the detection of SQL injection when the injection point is an URL (GET), and the third case, it is the SQL injection detection in authentication forms (POST).

VII.   THE DISTANCE USED

To analyze the similarity between two HTML pages, we need a distance to evaluate the deference between two strings. The order of words in page has a great importance. Thus, to calculate the distance between two pages we have selected Levenshtein algorithm [5]. The principle of this algorithm is described in figure 7.

The final distance is calculate as follow, and the result will be between [0,1] :

$$Dis = \frac{DistanceOfLevenshtein(Chaine1, Chaine2)}{lengthChaine1 + lengthChaine2}$$

VIII.   RESULTS AND EXPERIMENTS

We conducted a series of experiments on six Web applications that we have developed, containing seven secure pages and other seven unsecured, in total, in order to illustrate the effectiveness of our algorithm.

The main objective of these experiments is to characterize the ability of the model to deal with SQL injection attack and specifically to evaluate its effectiveness for detection. Such efficiency is generally characterized by the evaluation of the detection rate, false negative and false positive rate.

Note:

- We say that two pages a and b are similar if the distance between them is closer to the 0, and they are not similar otherwise.

- All you need is a single valid query that successfully exploits the flaw to say that the page is vulnerable to a SQL injection attack.

- The Model contains the following cardinality:

Cardinality Inval: 10 queries

Cardinality Aleat: 10 queries

Cardinality Val : 42 queries

In what follows, we will present the curve of each case.

A.   *A non-secure Web page*

   a) URL non secured

For a non-secure web page (URL), we have the following curves:

Fig. 8.   Curve of the distance between invalid and random

Fig. 9.   Curve of the distance between reference and random

Fig. 10. Curve of the distance between reference and invalid

For a non-secure Web page (URL), we have the following geometric representation:

Fig. 11. Curve of the result of a non-secure page (URL)

Figure 12 shows the distance between the three classes (Ref, Inval, Aleat) described in the green square. Inside each circle, we have the green dot that expresses mathematical average written in Pink Square and the mathematical variance written in the yellow square.

The distance between the classes is very high which means that the decoupling between them is very strong, the variance of each class is very small, which means that the values of each class are identical.

a)   *Unsecured formulary*

For a non-secure formulary, we have the following curves:
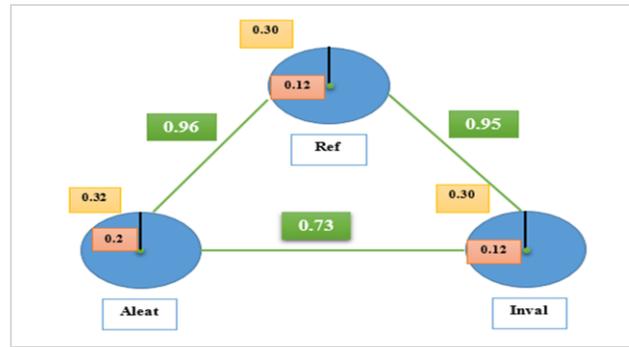
Fig. 12. Geometric representation (triangular) of a non-secure page (URL)
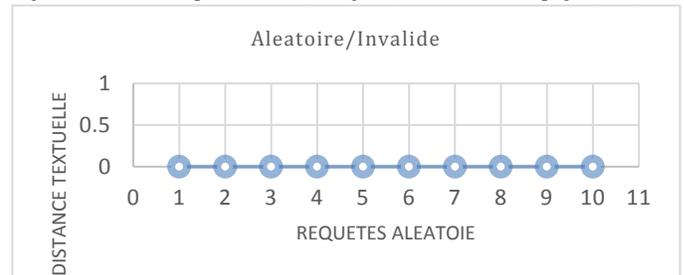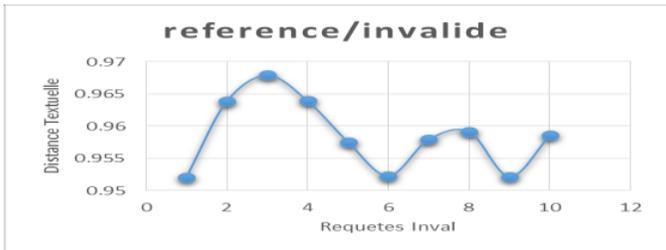
Fig. 13. Curve of the distance between random and invalide
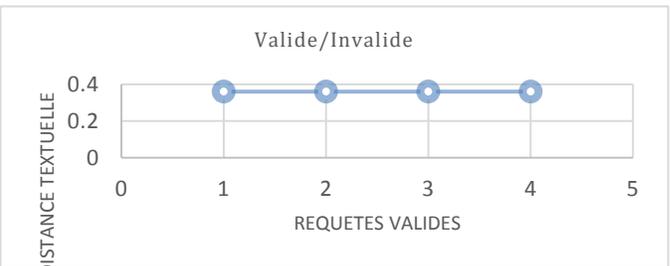
Fig. 14. Curve of the distance between valid and invalid

Fig. 15. Curve of the distance between valid and random
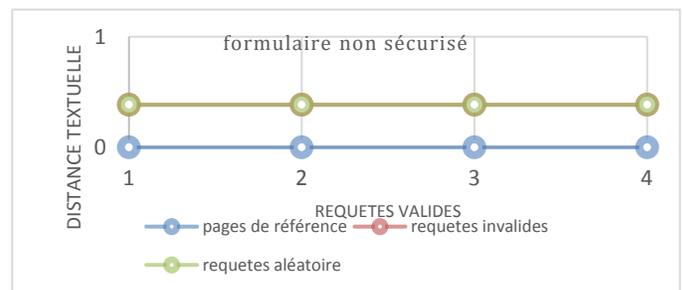
Fig. 16. Curve of the result of a non-secure page (formulary)

Fig. 17. Geometric representation (line) of a non-secure page (formulary)

In figure 13, we see that the distance of Inval/Aleat is 0 which is our algorithm singularity point (authentication form).

In figure 14 and 15, we see that the distance Val/Aleat, Val/Inval do not equal 0, this means that the field texts in this authentication formulary are vulnerable. The distance calculated in figure 16 prove that.

For this non-secure formulary, we have a geometries representation showed in figure 17.

B. A secure Web page

a) Secured URL

For a secure web page (URL), we have the following curves:
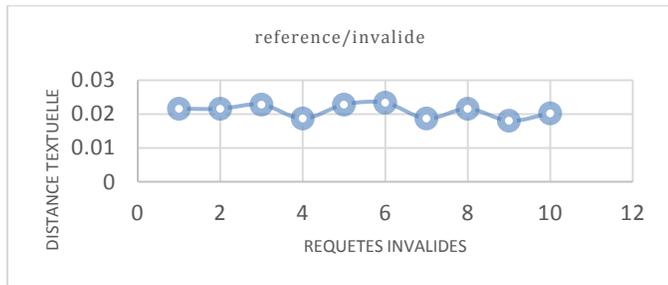
Fig. 18. Curve of the distance between reference and invalid
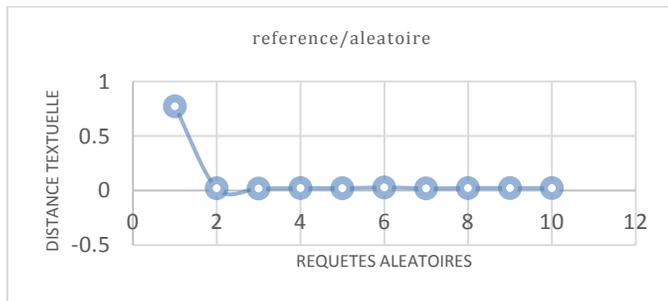
Fig. 19. Curve of the distance between reference and random
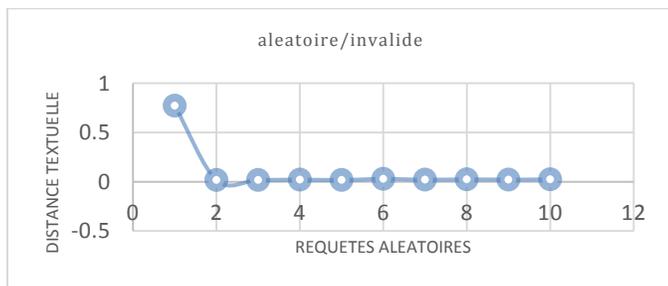
Fig. 20. Curve of the distance between random and invalid

Figures 18,19,20 and 21 shows the distance between all classes, this distance is near 0. This express that all classes are approached one to another wich mean that this unity lead to the security of Web page.

b) Secured authentication form
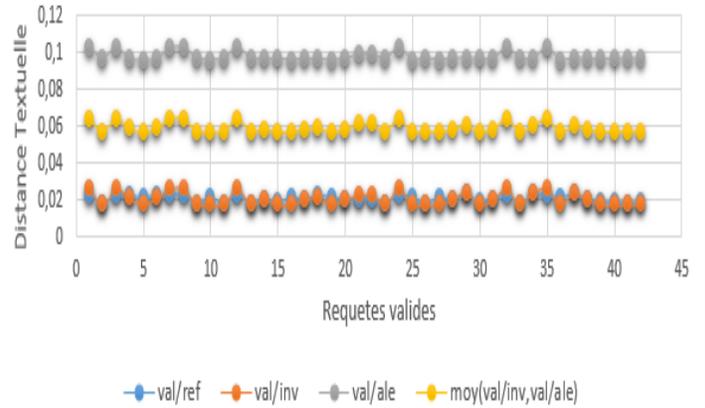

Fig. 21. Curve of the result of secure Web page (URL)
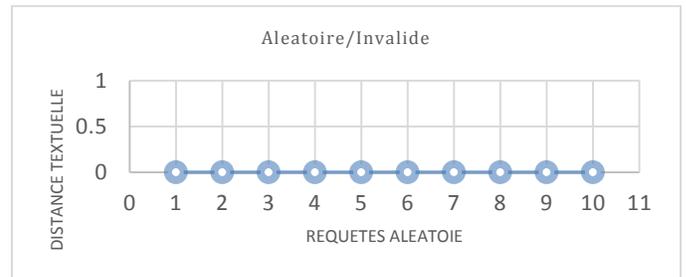
Fig. 22. Curve of the distance between random and invalid
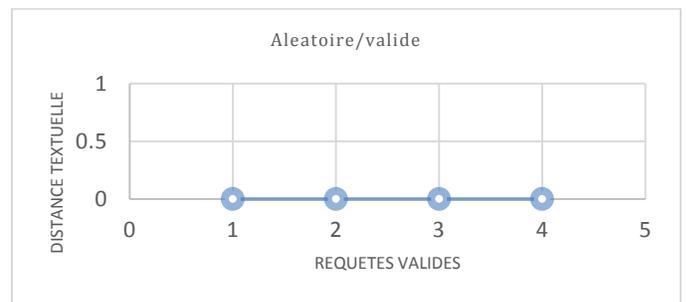

Fig. 23. Curve of the distance between random and valid
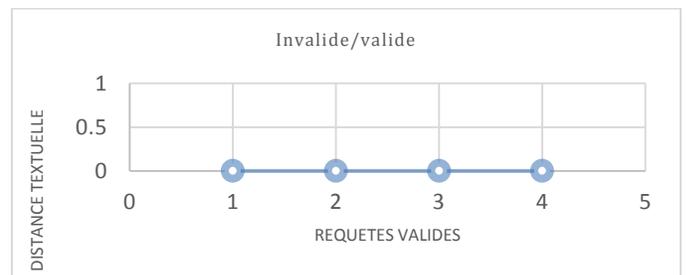
Fig. 24. Curve of the distance between invalid and valid

Fig. 25. Geometric representation (point) of a secure formulary

We apply our algorithm and the one described in the WASAPY model on Web application secure and non-secure, we had the following results (for a non-secure Web page URL):

TABLE I.   WASAPY AND LS MODEL RESULT

Valid queries		WASAPY	LS	
		$\varepsilon=0.75$	IS=0.73	x=0.955
	R_{15}, R_{19}, R_{20}, R_{26}, R30, R_{32}, R_{33}, R_{35}	✓	✓	
	R_3, R_{18}, R_{22}, R_{25}	✗	✓	
Number of detection		8	12	
Number of false positive		0	0	
Number of false negative		4	0	

For this page, we see that our algorithm detects all requests that must be detected with a 0 number of false positive and false negative while WASAPY model does not detect certain queries.

And for each non-secure Web application, we had the same result described by the graph below:
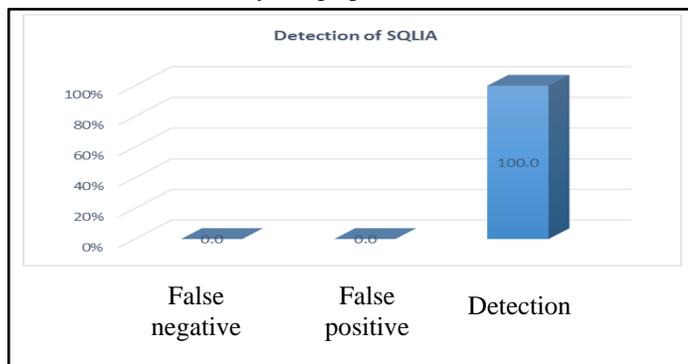
Fig. 26. Curve of the result of a non-secure Web application

IX.   CONCLUSION

We used the mathematical variance characterize the dispersion of our pattern. it shows how the statistical series or random variable is dispersed around its average. A variance of zero indicates that all values are identical. A small variance is a sign that the values are close to each other while a high variance is a sign that they are very open. We used the notion of variance to see the coupling factor of a class.

The idea of adding a fourth class that contains only legitimate pages for SQL injection detection in web applications has been successful which is proved by the results obtained and discussed.

The algorithm has a detection performance so interesting that it is wise to conduct an intensive study to confirm its validity, especially since it can detect all potential applications that leverage SQL injection. It should be noted that the model specifically detects requests injection successful that WASAPY approach don't detect. And so, the proposed Model LS, is a good step in website securing field.

REFERENCES

[1]   Sectools, http://sectools.org/web-scanners.html, consulté le 20/05/2013

[2]   Sourceforge, http://w3af.sourceforge.net, consulté le 22/04/2013.

[3]   Google, http://code.google.com/p/skipfish, consulté le 02/05/2013.

[4]   OWASP, www.owasp.org, accessed in 05/05/2013.

[5]   V. Levenshtein, Binary codes capable of correcting deletions, insertions and reversals", Soviet Physics Doklady, 1966, pages 707-710.

[6]   R. Akrout, Analyse de Vulnérabilité et Evaluation de Système de Détection d'Intrusion pour les Applications Web, Doctorat, Université de Toulouse, 2012.

[7]   R. Akrout, Analyse de Vulnérabilité et Evaluation de Système de Détection d'Intrusion pour les Applications Web, Doctorat, Université de Toulouse, 2012, pages 53-54.

Introducing Concurrency to Workflows: Theory and A Real-World Case Study

Laird Burns, Wai Yin Mok, and Wes N. Colley

College of Business Administration
University of Alabama in Huntsville
Huntsville, Alabama, USA, 35899

Abstract—**Making use of the concepts of the activity diagrams of the Unified Modeling Language, this paper defines an important class of workflows, called flow independent workflows, which are deterministic in the sense that if a flow independent workflow is given the same multi-set of resources as input over and over again, it will produce the same output every time. After which, this paper provides a methodology, and its accompanying algorithms, that introduces concurrency to flow independent workflows by rearranging the action nodes in every flow of control. It then applies the methodology to a real-world case to demonstrate its usefulness. It concludes with future research possibilities that might extend the methodology to general workflows, i.e., not necessarily flow independent workflows.**

Keywords—*Business process re-engineering; Unified Modeling Language; activity diagrams; flow independent workflows; concurrency*

I. INTRODUCTION

Business process re-engineering (BPR) has been a subject of intensive study since Michael Hammer published his seminal paper "Reengineering Work: Don't Automate, Obliterate" [6]. As a business management strategy, it focuses on the analysis and design of workflows and processes within an organization. Improving customer service and reducing operational costs are among its many goals. Many times BPR involves large scale modifications and redesigns of existing workflows. This paper, however, introduces a methodology of smaller scale that can speed up a specific class of workflows that has certain desirable properties. We then supplement the methodology with a real-world case study that demonstrates the usefulness of the methodology.

To formally prove the methodology really does what it claims, we need a mathematical foundation upon which the methodology is based. For this purpose, we adopt some of the ideas of The Unified Modeling Language (UML), which has been an industry standard for modeling software-intensive systems since 2000 [12]. Developed by Grady Booch, Ivar Jacobson and James Rumbaugh at Rational Software in the 1990s, UML provides a set of graphic notation techniques to create visual models of object-oriented software-intensive systems [3]. Particularly relevant to this research is the activity diagrams, from which our methodology is derived.

Another important component of our methodology is the database concurrency theory [2], which has been an area of study in the last thirty years. Many recent studies applied this theory to workflows [1,4,5,7,8,9,10,11]. Database concurrency theory is mainly concerned with maintaining an orderly access of data items in the presence of multiple long-running concurrent database transactions. Without proper control, concurrent database transactions might read from and write to data items in an arbitrary order that will take a database from a correct state to an incorrect state. Placing locks on data items at the correct moment is a main mechanism the theory uses to control the execution of the transactions to ensure that the database will go from one correct state to another correct state.

This paper is organized as follows. In Section 2, we present a mathematical foundation for workflows and define a particular class of workflows, called flow independent workflows, that has the property that being deterministic. A workflow being deterministic means that it will always produce the same output for a given input. Section 3 presents the proposed workflow re-engineering methodology and the correctness proof. Section 4 gives the case study. We conclude and point out possible future researches in Section 5.

II. FUNDAMENTALS

A. Workflows and UML Activity Diagrams

UML has two types of diagrams: structural diagrams and behavioral diagrams [3]. Among all the behavioral diagrams of UML, activity diagrams are particularly relevant to this research because they specify the operational step-by-step processes of a system. Many concepts of activity diagrams are especially useful. For example, our definition for workflows adopts branches of control and concurrency from activity diagrams. Further, the concept of resources is also necessary in the definition because many workflows use, modify, consume and produce resources as they are executed. In addition, we also define the rules that govern the execution of workflows, i.e., the execution semantics of workflows.

Definition 1: A *workflow* is an 8-tuple (*Actions*, *Branches*, *Merges*, *Forks*, *Joins*, *Arrows, Resources*, *Constraints*). *Actions*, *Branches*, *Merges*, *Forks*, *Joins* are all finite sets of nodes (vertices) and *Arrows* is a finite set of directed edges (ordered pairs) of nodes in *Actions* ∪ *Branches* ∪ *Merges* ∪ *Forks* ∪ *Joins*. *Resources* is a finite multi-set of resources and *Constraints* is a finite set of constraints over the workflow. The following rules further refine these concepts.

1: *Resources* is a finite multi-set of $n \geq 1$ resources r_1, r_2, …, r_n. A resource r_i of *Resources* can only be in one of an enumerable number of states at any moment of the execution of the workflow. (A workflow might need more than one unit

of a resource type in its execution. Therefore, a multi-set, rather than a set, of resources is used in the definition because elements of a multi-set are not necessarily distinct. However, the subscripts can make the elements distinct. For example, two identical but distinct elements r_i and r_j in *Resources* can be distinguished by their subscripts if $i \neq j$. In a sense, the subscripts tag the resources and serve as unique identifiers.)

2: An action node in *Actions* denotes an action, which is executed completely or not at all. An action takes a finite amount of time to complete, although it may vary from time to time. Any action node v_i in *Actions* yields at most one directed edge of the form (v_i, v_j) in *Arrows*. (In other words, each action node has at most one out arrow.)

3: Each action node has four associated multi-sets of resources: *use*, *modify*, *consume*, and *produce*. *Use* is the finite multi-set of resources that the action uses in its execution. *Modify* is the finite multi-set of resources whose states will be changed by the action during its execution. *Consume* specifies the finite multi-set of resources that the action consumes during its execution. After the action consumes a resource r_i, r_i ceases to exist. *Produce* is the finite multi-set of resources that the action produces during its execution. The action can assign a resource r_i in one of its many allowable states when the action produces r_i. Naturally, *use* $\supseteq$ *modify* $\cup$ *consume* and *use* $\cap$ *produce* $= \emptyset$. We stipulate that if an action node is given the same multi-set *use* multiple times, the action node will always yield the same multi-sets *modify*, *consume*, and *produce* every time.

4: The nodes in *Branches* $\cup$ *Merges* $\cup$ *Forks* $\cup$ *Joins* are *control nodes*. Control nodes only direct the flows of control during the execution of the workflow and thus they do not modify, consume or produce the resources in *Resources*.

5: A *source node* n is a node that there is not a directed edge (v_i, n) in *Arrows*. A *sink node* n is a node that there is not a directed edge (n, v_i) in *Arrows*. Further, none of the directed edges in *Arrows* has the form (n, n) where $n \in$ *Actions* $\cup$ *Branches* $\cup$ *Merges* $\cup$ *Forks* $\cup$ *Joins*. (In other words, self-loops are not allowed.) Source nodes and sink nodes must be action nodes, although multiple source nodes and/or sink nodes are allowed.

6: Initially, there is not a flow of control in the workflow. When the execution of the workflow begins, simultaneously a separate flow of control will start at each source node. The execution of the workflow terminates when all of its flows of control terminate.

7: When a flow of control reaches v_i of a directed edge (v_i, v_j), whether v_i will start executing depends on if v_i uses any resources. If v_i does not use any resources, v_i immediately starts executing. If v_i uses some resources, then we further check whether the elements in v_i's *use* are also in *Resources*. If *use* $\subseteq$ *Resources*, then v_i starts executing. In both cases, after v_i stops executing, *Resources* = *Resources* $\cup$ v_i's *produce* $-$ v_i's *consume*, the elements in v_i's *modify*, which are also in *Resources*, have been modified, and v_j immediately starts executing. If *use* $\nsubseteq$ *Resources*, then the flow of control terminates at v_i with no change to *Resources*. (In the case that the resources of v_i are not immediately available to v_i when the

flow of control reaches v_i, v_i may enter a waiting period, which can be modeled by adding a branch node of Rule 8 and/or a constraint of Rule 12. Example 3 illustrates how this can be done. Hence, there is no loss of generality to say that if *use* $\nsubseteq$ *Resources*, then the flow of control terminates at v_i. Note that any action of a reasonable workflow will not wait indefinitely. Therefore, we stipulate that the length of such a waiting period must be specified beforehand.)

The following rules are concerned with branch nodes, merge nodes, fork nodes and join nodes. All of these nodes are not source nodes and sink nodes. As a result, for every node v_j $\in$ *Branches* $\cup$ *Merges* $\cup$ *Forks* $\cup$ *Joins*, there are nodes v_i and v_k such that (v_i, v_j) and (v_j, v_k) are both in *Arrows*.

8: For any branch node v_j in *Branches*, there is a unique directed edge (v_i, v_j) in *Arrows* and there are $n \geq 1$ directed edges $(v_j, v_{k1}), (v_j, v_{k2}), \ldots, (v_j, v_{kn})$ in *Arrows*. Further, there is a condition associated with (v_j, v_{kl}), $1 \leq l \leq n$. When a flow of control reaches v_j, at most one condition out of those associated with these n directed edges can be true at that moment. The flow of control then follows the unique directed edge whose condition is true. If the conditions associated with these n directed edges are all false, then the flow of control terminates at v_j. Since the conditions associated with the directed edges $(v_j, v_{k1}), (v_j, v_{k2}), \ldots, (v_j, v_{kn})$ might involve the elements in *Resources*, v_j also has a multi-set *use*, which specifies the elements in *Resources* that are used in these conditions.

9: For any merge node v_j in *Merges*, there is a unique directed edge (v_j, v_k) in *Arrows* and there are $n > 1$ directed edges $(v_{i1}, v_j), (v_{i2}, v_j), \ldots, (v_{in}, v_j)$ in *Arrows*. For any flow of control that reaches any of $v_{i1}, v_{i2}, \ldots, v_{in}$ and after it stops executing, v_k immediately starts executing. Note that v_j does not need any resources in its execution.

10: For any fork node v_j in *Forks*, there is a unique directed edge (v_i, v_j) in *Arrows* and there are $n > 1$ directed edges $(v_j, v_{k1}), (v_j, v_{k2}), \ldots, (v_j, v_{kn})$ in *Arrows*. When a flow of control reaches v_j via v_i, the flow of control terminates at v_j and n new flows of control $f_1, f_2, \ldots, f_n$ will be created, and each f_l starts at v_j and then immediately reaches v_{kl}, $1 \leq l \leq n$. Note that v_j does not need any resources in its execution.

11: For any join node v_j in *Joins*, there is a unique directed edge (v_j, v_k) in *Arrows* and there are $n > 1$ directed edges $(v_{i1}, v_j), (v_{i2}, v_j), \ldots, (v_{in}, v_j)$ in *Arrows*. A new flow of control starts at v_j implies that there are n flows of control $f_1, f_2, \ldots, f_n$ and each f_l, $1 \leq l \leq n$, reaches v_j via v_{il} and then terminates at v_j. We stipulate that every flow of control that reaches v_j via v_{il} for some l can only give raise to a single new flow of control starting at v_j. Note that v_j does not need any resources in its execution.

12: *Constraints* is a finite set of constraints defined on the workflow. $\square$

Example 1: A moving company workflow might require a multi-set {truck$_1$, truck$_2$, worker$_1$, worker$_2$, worker$_3$, worker$_4$} of two trucks and four workers to move a family. Although truck$_1$ and truck$_2$ might be identical, the subscripts distinguish them from each other. As for the status of a resource, for example a document might be in one of the states prepared,
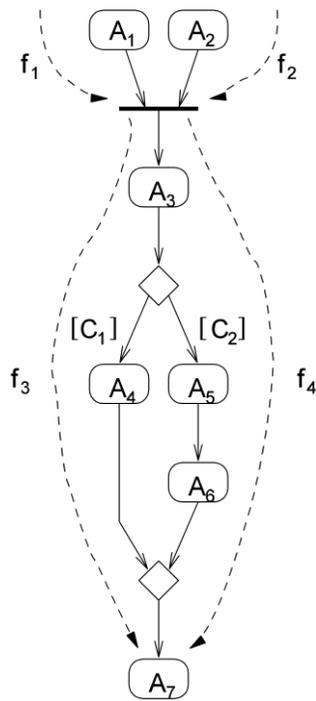
Fig.1.   A sample workflow in UML notation

unsigned, and signed; and a temperature variable might assume one of its many permissible values. □

In most workflows, many actions cannot have a fixed duration to complete. For example, a mailman might be able to finish his job in 5 hours on a sunny day but not on a rainy day. Hence, unless otherwise specified, we leave the duration of an action open. When such constraints are necessary, they are added to the set *Constraints*.

We use UML notation to represent the components of a workflow. Action nodes are represented by rounded boxes, branch nodes and merge nodes by diamonds and fork nodes and join nodes by thick lines. Arrows naturally represent directed edges.

Example 2: Fig 1shows a workflow in UML notation. When the workflow begins executing, two flows of control f_1 and f_2, one starting at A_1 and the other at A_2, start simultaneously. These two flows of control are executed concurrently and synchronized at the join node, where each waits until the other reaches the join node. Then, they both terminate at the join node and a new flow of control starts at the join node. At the branch node, either C_1 or C_2 but not both is true. The flow of control follows the directed edge whose condition is true. Therefore, there are two possible flows of control f_3 and f_4 starting at the join node; but unlike f_1 and f_2, f_3 and f_4 cannot co-exist at the same time. They are merely two different possibilities. Finally, the flow of control reaches the merge node, where it simply continues on to A_7, the final action node. Although not shown in the Fig., the workflow in has a multi-set *Resources* and each action node has four associated

finite multi-sets *use*, *modify*, *consume* and *produce*. □
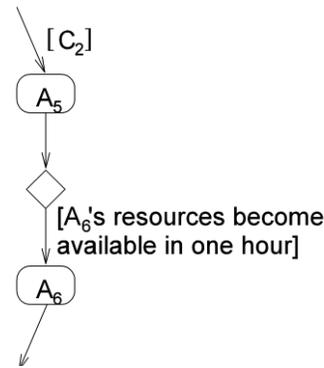
Fig.2.   Adding a waiting period to A6 in Fig.1.

Example 3: Suppose the resources of action node A6 in are not available immediately when a flow of control reaches A6. A6 then has to wait for a while. To capture such a waiting period, a branch node can be added before A6 in the workflow. Since workflows cannot wait forever, there must be a specific time limit for the waiting period. Hence, the condition in specifies that the waiting period must be less than one hour, although some other time limit is also possible. If A6's resources become available within one hour, the follow of control continues to A6; otherwise, the condition fails and the flow of control terminates at the branch node. □

Example 4: If initially *Resources* = {a, b, c_1, c_2} and A_1's *use* = {a, d_1, d_2}, then the flow of control started at A_1 immediately terminates at A_1 because A_1 uses two *d*'s in its execution, but there is none available in *Resources*. When that happens, even if A_2 successfully completes its action, no new flow of control will start at the join node because the join node will indefinitely wait for a flow of control coming from A_1, which will never come. □

Example 5: We might additionally add a constraint that specifies A_1 must complete its action within an hour or that the total time the workflow will take must be less than 5 hours. These constraints will then be added to the set *Constraints*. □

B. *Configuartions of Workflows*

Although self-loops are not allowed in a workflow, cycles are still possible. Hence, a workflow might not stop executing once it starts. Since most useful workflows terminate once they are given enough time and resources, non-terminating workflows are not considered any further.

Definition 2: A *terminating workflow* is one that will terminate on any given multi-set *Resources*. □

Example 6: Because it does not have any cycles, the workflow in Fig.1 always terminates not matter what the multi-set *Resources* is. □

Definition 3: A *configuration* of a workflow is the multi-set *Resources* together with the states of the elements in *Resources*. □

Every workflow has an initial configuration, which is the multi-set *Resources* together with the states of each of its elements before the workflow starts executing. After the

workflow stops executing, it then has a final configuration, which is the multi-set *Resources* together with the states of each of its elements after the workflow stops executing.

Definition 4: A *deterministic workflow* is a terminating workflow and for any given multi-set *Resources*, it always ends up in the same final configuration even if the workflow has to be executed multiple times. On the other hand, a *non-deterministic workflow* is also a terminating workflow and for some given multi-set *Resources*, it might not end up in the same final configuration if the workflow is executed multiple times. □

Although a deterministic workflow is a terminating workflow, the converse is not true. That is, it is possible that a terminating workflow is non-deterministic. Since most real-world workflows are supposed to give the same results even if they have to be executed more than once, deterministic workflows are desirable.

Example 7: Suppose *Resources* = {a, b_1, b_2, c}, and A_1's *use* = A_1's *modify* = A_2's *use* = A_2's *modify* = {a}, and in addition resource a is not in the multi-sets *modify*, *consume* and *produce* of the other action nodes in Fig.1. Since A_1 might finish before A_2 in one run, but in another run A_2 might finish before A_1, and since A_1 and A_2 both modify resource a, the workflow in Fig.1 would be non-deterministic under these conditions. □

Here, we give a sufficient condition for deterministic workflows. That is, if a workflow satisfies this condition, the workflow is deterministic.

Definition 5: Two (not necessarily distinct) actions A_i and A_j are *independent* if A_i's *use* ∩ (A_j's *modify* ∪ A_j's *consume* ∪ A_j's *produce*) = ∅ and A_j's *use* ∩ (A_i's *modify* ∪ A_i's *consume* ∪ A_i's *produce*) = ∅. □

Definition 6: Two distinct concurrent flows of control f_i and f_j of a workflow are *independent* if any pair of actions A_i of f_i and A_j of f_j are independent. A workflow is *flow independent* if any two of its distinct concurrent flows of control f_i and f_j are *independent*. □

Example 8: f_1 and f_2 are distinct concurrent flows of control in Fig.1. f_3 and f_4 are not concurrent flows of control because they cannot co-exist at the same time. □

Theorem 1: If a workflow w is flow independent, w is deterministic.

Proof sketch: We proceed by induction on the number n of concurrent flows of control of w at any moment. If $n = 1$, then at any moment w has at most one flow of control. Suppose this flow of control f_i has $n ≥ 1$ nodes v_1, v_2, …, v_{n-1}, v_n. By Rule 3 of Definition 1, when an action node is given the same multi-set *use* multiple times, the action node will always yield the same multi-sets *modify*, *consume*, and *produce* every time. Hence, every action node v_i of f_i is deterministic. It is clear that

a sequential execution of $n ≥ 1$ deterministic action nodes is also deterministic. Hence, the basis is established. Assume the induction hypothesis is true for $n = k$ for some $k ≥ 1$. That is, w is deterministic if w has k or less concurrent flows of control at any moment. Suppose w has one more flow of control f_{k+1} at a moment. By the assumption that w is flow independent, any pair of actions A_i of f_{k+1} and A_j of f_l ($1 ≤ l ≤ k$) are independent. Hence, every action A_i of f_{k+1} does not depend on the output of the action nodes of f_l ($1 ≤ l ≤ k$). At this point the argument for the basis also applies to f_{k+1}. Hence, the workflow is deterministic. □

III. WORKFLOWS RE-ENGINEERING

The main purpose of workflows re-engineering is to make workflows more efficient, or to reduce certain resources required by the workflows. For most cases, the goal is to reduce the completion time of a workflow. Because flow independent workflows have the desirable property that they are deterministic, in this section we focus on flow independent workflows and present a methodology that will reduce their completion time.

A. Introducing Concurrency

Introducing concurrency to a flow independent workflow obviously can reduce the time it takes to complete. However, the interactions of the actions and the resources of the workflow imply a certain order the actions must observe. Example 9: Consider the workflow in and suppose A_5's *use* = {a} and A_6's *modify* = {a}. Under this assumption, A_5 uses the resource a before it is modified by A_6. Hence, A_5 must precede A_6. Now suppose A_5's *modify* = {a} and A_6's *use* = {a}. In this case, A_6 uses the resource a modified by A_5. Hence, A_6 must follow A_5. Similarly, suppose A_5's *use* = {a} and A_6's *consume* = {a}. Then, A_5 must precede A_6 because after A_6, resource a will cease to exist. If A_5's *consume* = {a} and A_6's *use* = {a}, then A_6 cannot be executed because resource a no longer exists after A_5. Hence, the flow of control terminates at A_5. (Note that one may argue that it is incorrect to have A_5's *consume* = {a} and A_6's *use* = {a}. However, we have to be compliant to the semantics of the workflow and thus the order of execution of the actions in the workflow must be preserved.) For the case that A_5's *produce* = {a} and A_6's *use* = {a}, A_5 must precede A_6. On the other hand, if A_5's *use* = {a} and A_6's *produce* = {a}, then the flow of control terminates at A_5, which is similar to the case that A_5's *consume* = {a} and A_6's *use* = {a}. However, if A_5 and A_6 are independent, they can be executed concurrently. This can be done by adding a fork node as shown in Fig 3 As a result, two concurrent flows of control can start at A_5 and A_6 simultaneously. □

The main idea of our methodology is to introduce concurrency to a flow independent workflow; but at the same time we have to ensure that the resulting workflow is equivalent to the original.

if A_5 and A_6 are independent
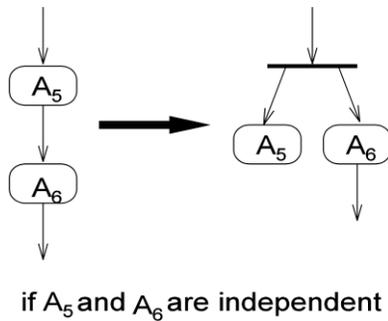
Fig.3. A_5 and A_6 can be executed concurrently if they are independent.

Definition 7: Two deterministic workflows are *equivalent* if they are given any multi-set *Resources* as their initial configurations, they will both end up in the same final configuration after they stop executing. □

B. *Partitioning Flows of Control into Sequences*

Lemma 1 points out some useful characteristics that will help introducing concurrency to flow independent workflows.

Lemma 1: Consider a flow of control f_i of $n \geq 1$ nodes v_1, v_2, …, v_{n-1}, v_n of a flow independent workflow. The following are all true for f_i.

a) v_1 cannot be a branch node.

b) Both v_1 and v_n cannot be a merge node.

c) Only v_1 or v_n can be a fork node.

d) Only v_1 or v_n can be a join node.

e) If v_i $(1 < i < n)$ is a branch node, then each of v_1, …, v_{i-1} cannot be executed concurrently with any of v_{i+1}, …, v_n.

Proof sketch:

a) Rule 8 of Definition 1 states that when a flow of control reaches a branch node, the conditions associated with all of its outgoing directed edges will be evaluated and the flow of control will follow the unique directed edge whose condition is true. Hence, branch nodes do not start any flow of control and thus v_1 cannot be a branch node.

b) According to Rule 9 of Definition 1, merge nodes neither start nor terminate any flow of control. Given a merge node v_j, there are two directed edges (v_i, v_j) and (v_j, v_k) in Arrows and v_j simply passes the flow of control from v_i to v_k. Hence, v_1 and v_n cannot be a merge node.

c) Rule 10 of Definition 1 states that when a flow of control reaches a fork node, the flow of control will terminate at the fork node and then the fork node will yield more than one flow of control. Hence, if a v_i where $1 < i < n$ is a fork node, then the flow of control starting at v_1 will terminate at v_i. This contradicts the lemma statement that the flow of control terminates at v_n.

d) Rule 11 of Definition 1 states that when a flow of control reaches a join node, the flow of control will wait for and synchronize with the other flows of control destined for the join node. Then, they will all terminate at the join node and a new flow of control will start at the join node. Hence, if a v_i where $1 < i < n$ is a join node, the flow of control starting at v_1 will terminate at v_i. This contradicts the lemma statement that the flow of control terminates at v_n.

e) If v_i where $1 < i < n$ is a branch node, then by Rule 8 of Definition 1 there is a condition associated with the directed edge (v_i, v_{i+1}). The execution of any of v_{i+1}, …, v_n depends on the truth value of the condition. If the condition is true, v_{i+1}, …, v_n will be executed in this order; otherwise, none of them will be executed. On the other hand, the execution of v_1, …, v_{i-1} do not depend on the condition of the directed edge (v_i, v_{i+1}). As a result, each of v_1, …, v_{i-1} cannot be executed concurrently with any of v_{i+1}, …, v_n. □

While the other parts of Lemma 1 restrict the nodes in f_i, Part e relates directly to concurrency. With the additional restriction that none of the nodes of f_i can be replicated, Lemma 2 provides another result about merge nodes that also relates directly to concurrency.

Lemma 2: Consider a flow of control f_i of $n \geq 1$ nodes v_1, v_2, …, v_{n-1}, v_n of a flow independent workflow. If v_j $(1 < j < n)$ is a merge node and none of the nodes of f_i can be replicated, then each of v_1, …, v_{j-1} cannot be executed concurrently with any of v_{j+1}, …, v_n.

Proof sketch: By Rule 9 of Definition 1, there are at least two directed edges (v_{j-1}, v_j) and (v_i, v_j) and a unique directed edge (v_j, v_{j+1}) for v_j. v_{j-1} is the node precedes v_j in f_i but v_i is not part of f_i. When a flow of control reaches v_j via v_{j-1}, v_j will pass the flow of control to v_{j+1}. Likewise, when a flow of control f_k $(i \neq k)$ reaches v_j via v_i, v_j will pass the flow of control to v_{j+1}. If any of v_{j+1}, …, v_n is executed concurrently with any of v_1, …, v_{j-1}, the same node must be removed from v_{j+1}, …, v_n and must also be executed concurrently with some node before v_j on f_k as well. However, since nodes on f_i cannot be replicated, this is impossible. □

Given a flow of control f_i of a flow independent workflow, Lemmas 1 and 2 specifies certain restrictions on f_i. Particularly, the branch nodes and merge nodes partition f_i into sequences of action nodes. By Lemma 1e and Lemma 2, every action node of each of these sequences cannot be executed concurrently with any action node of the other sequences of f_i. However, we can rearrange the action nodes within a sequence so that some of them can be executed concurrently.

Example 10: Consider a flow of control f_i of a flow independent workflow that has two diamonds, where each diamond either represents a branch node or a merge node. As shown in Fig 4 these two diamonds partition f_i into three sequences of action nodes, where each sequence of action nodes is represented by a dashed line. □

To rearrange the action nodes within a sequence, we first need to define a relation on the action nodes. This relation will yield an order of execution on the action nodes.

Definition 8: Given a sequence of $n \geq 1$ action nodes v_1, v_2, …, v_{n-1}, v_n, we define a relation, denoted by $\lhd$, on the action nodes as follows:
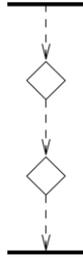
Fig.4. Three sequences of action nodes as a result of two branch/merge nodes.

a) $v_i \lhd v_j$ if $i < j$ and v_i and v_j are not independent, or

b) $v_i \lhd v_j$ if $v_i \lhd v_k$ and $v_k \lhd v_j$ for some v_k. □

The meaning of $v_i \lhd v_j$ is that v_i must be executed before v_j. Condition a is obvious because if v_i precedes v_j in the sequence and they are not independent, then v_i must be executed before v_j. (See Example 9.) Condition b says that if v_i must be executed before v_k and v_k must be executed before v_j, then v_i must (indirectly) be executed before v_j. Based on the relation $\lhd$, we can now define an execution order hierarchy for the action nodes as follows:

Definition 9: Given a sequence of $n \geq 1$ action nodes v_1, v_2, …, v_{n-1}, v_n, we define an execution order hierarchy h for the action nodes as follows:

a) If $v_i \not\lhd v_k$ for each v_i such that $v_i \neq v_k$, then v_k is a root of h. (Note that it is possible that h may have more than one root.)

b) If v_i is a node in h and $v_i \lhd v_j$ and there does not exist an v_k such that $v_i \lhd v_k$ and $v_k \lhd v_j$, then v_i is a parent of v_j in h. (Note that an action node may have more than one parent in h. Also note that if $v_i \lhd v_k$ and $v_k \lhd v_j$ for some v_k, then v_i is not a parent of v_j in h. However, v_i is an ancestor of v_j in h.)

Algorithms can be easily defined to generate the execution order hierarchy of Definition 9. One possibility is as follows.

Algorithm 1:

Input: a sequence of $n \geq 1$ action nodes v_1, v_2, …, v_{n-1}, v_n.

Output: an initially empty execution order hierarchy h of Definition 9 for the input action nodes.

1) Set $s = \emptyset$.

2) For $i = 1$ to n do

For $j = i+1$ to n do /* $i < j$ */

If v_i and v_j are not independent, insert (v_i, v_j) to s.

3) For $i = 1$ to n do

If v_i does not appear in h, make v_i a root of h.

For each ordered pair of the form (v_i, v_j) in s,

add a directed edge $v_i \rightarrow v_j$ to h.

4) For $i = 1$ to n do /* Remove redundant edges from h. */

If $v_i \rightarrow v_k$ is a directed edge in h but $v_i \rightarrow v_k$ can be derived from the other directed edges in h, remove $v_i \rightarrow v_k$ from h.

Example 11: Consider the sequence of action nodes in Fig 5 and further suppose the following:

1) A_1's *use* = $\{a\}$, A_1's *modify* = $\{a\}$, A_1's *produce* = $\{b, c\}$,

2) A_2's *use* = $\{d\}$, A_2's *produce* = $\{e, f\}$,

3) A_3's *use* = $\{a, b, e\}$, A_3's *modify* = $\{a\}$, A_3's *produce* = $\{g\}$,

4) A_4's *use* = $\{c, f\}$, A_4's *modify* = $\{f\}$, A_4's *consume* = $\{c\}$,

5) A_5's *use* = $\{a\}$, A_5's *produce* = $\{h\}$,

6) A_6's *use* = $\{e\}$, A_6's *produce* = $\{i\}$.
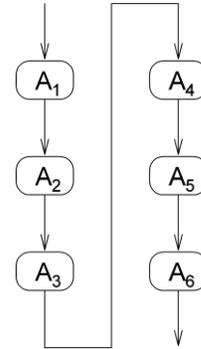
Fig.5. A sequence of action nodes
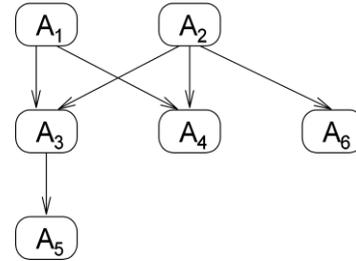
Fig.6. The execution order hierarchy of Example 11

After Step 2 of Algorithm 1, $s = \{A_1 \rightarrow A_3, A_1 \rightarrow A_4, A_1 \rightarrow A_5, A_2 \rightarrow A_3, A_2 \rightarrow A_4, A_2 \rightarrow A_6, A_3 \rightarrow A_5\}$. In the first iteration of the for-loop of Step 3, A_1 becomes the first root of h. Then, the directed edges $A_1 \rightarrow A_3, A_1 \rightarrow A_4, A_1 \rightarrow A_5$ are added to h. In the second iteration of the for-loop of Step 3, A_2 becomes the second root of h. Then, the directed edges $A_2 \rightarrow A_3, A_2 \rightarrow A_4, A_2 \rightarrow A_6$ are added to h. In the third iteration of the for-loop of Step 3, A_3 does not become another root of h because A_3 already appears in h. Then, the directed edge $A_3 \rightarrow A_5$ is added to h. Now, the directed edge $A_1 \rightarrow A_5$ becomes redundant in h because $A_1 \rightarrow A_5$ can be derived from $A_1 \rightarrow A_3$ and $A_3 \rightarrow A_5$, which are also in h. Thus, $A_1 \rightarrow A_5$ is removed at Step 4. The result is the execution order hierarchy in Fig 6. □

Theorem 2: Given a sequence of $n \geq 1$ action nodes v_1, v_2, …, v_{n-1}, v_n, Algorithm 1 generates the execution order hierarchy of Definition 9 from the action nodes.

Proof sketch:

Step 2 of Algorithm 1 generates all ordered pairs $v_i \lhd v_j$ where there is no v_k such that $v_i \lhd v_k$ and $v_k \lhd v_j$. (That is, Step 2 of Algorithm 1 generates all ordered pairs $v_i \lhd v_j$ such that v_i must be executed directly, not indirectly, before v_j.) In fact, Step 2 of Algorithm 1 may as well generate some ordered pairs that can be derived from the other ordered pairs. Step 3 of Algorithm 1 simply adds all the directed edges to h based on the ordered pairs in s. Potentially it may add redundant directed edges to h that can be inferred from the other directed edges, e.g., $A_1 \longrightarrow A_5$ of Example 11. Step 4 of Algorithm 1 is designed to remove them. □

Transforming the execution order hierarchy of Definition 9 into a workflow with concurrency is straightforward. The idea is to use fork nodes to fork multiple flows of control and join nodes to synchronize these flows of control back into one.

Algorithm 2:

Input: an execution order hierarchy h generated by Algorithm 1.

Output: a workflow that is equivalent to the input sequence of action nodes to Algorithm 1.

1) For each action node v_i in h do,
If v_i has $n > 1$ outgoing directed edges, then

Add a new fork node f and a directed edge from v_i to f.

Make f to have n outgoing directed edges.

2) For each action node v_i in h do,

If v_i has $n > 1$ incoming directed edges, then

Add a new join node j and a directed edge from j to v_i.

Make j to have n incoming directed edges.

3) According to the directed edges in h, connect the added fork nodes and the join nodes and the rest of the action nodes in h.

4) If h has more than one root, add a fork node at the beginning of the resulting workflow such that it has out-going directed edges to all of the roots.

5) If h has more than one sink action node, add a join node at the end of the resulting workflow such that it has in-coming directed edges from all of the sink action nodes.

Example 12: In Fig 6, A_1 has two out-going directed edges. Thus, Algorithm 2 adds a fork node that has two out-going directed edges and a directed edge from A_1 to that fork node. Algorithm 2 does the same to A_2. A_3 has two in-coming directed edges. Thus, Algorithm 2 adds a join node that has two in-coming directed edges and a directed edge from that join node to A_3. Similarly, Algorithm 2 does the same to A_4. According to the directed edges in h, Algorithm 2 then connects the added fork nodes to the added join nodes and also the rest of the action nodes by directed edges. Because the execution order hierarchy has two roots, Algorithm 2 adds a fork node at the beginning of the workflow with out-going directed edges to the two roots. Since there are three sink action nodes, Algorithm 2 adds a join node at the end of the

workflow to merge all of the flows of control back into one. The resulting workflow is shown in Fig 7. □

Theorem 3: Given a sequence of $n \geq 1$ action nodes v_1, v_2, …, v_{n-1}, v_n, our methodology generates an equivalent workflow from the sequence of action nodes.

Proof sketch:

The resulting workflow in UML notation is a simple translation from the execution order hierarchy of Definition 9. By Theorem 2, Algorithm 1 generates the execution order hierarchy of Definition 9 from the sequence of action nodes. It is therefore sufficient to consider the execution order hierarchy generated by Algorithm 1. We proceed by induction on the number n of action nodes in the sequence. When $n = 1$, it is clear that the sequence of action nodes is equivalent to the
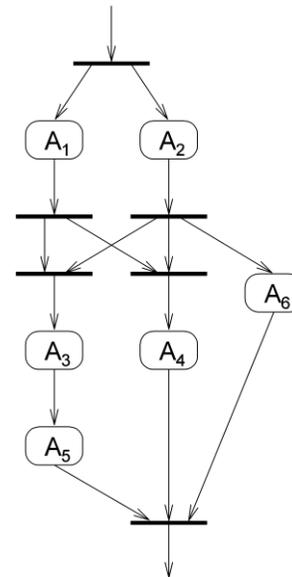
Fig.7. The resulting workflow generated from the execution order hierarchy in Fig 6.

execution order hierarchy because both of them have only one action node. We assume the induction hypothesis is true for $n = k$ for some $k \geq 1$. When $n = k+1$, the sequence has $k+1$ action nodes where v_{k+1} is the last action node added to the sequence. Step 2 of Algorithm 1 adds the ordered pair (v_i, v_{k+1}) to s for each $i \leq k$ such that v_i and v_{k+1} are not independent. Step 3 of Algorithm 1 adds the directed edge $v_i \longrightarrow v_{k+1}$ to the execution order hierarchy. As such, it is clear that if there is not a direct path from one action node v_i to another action node v_j in the execution order hierarchy, they are independent and thus v_i can be executed before v_j, or v_j can be executed before v_i, or v_i and v_j can be executed concurrently. Hence, the generated execution order hierarchy is equivalent to the original sequence of action nodes. □

IV. CASE STUDY

The case study is drawn from a representative workflow ("phase" in research client terminology) from a much larger set of workflows at a large organization which serves as a broker between major international customers and multiple US

suppliers ("Broker"). Broker provides a wide range of products and services to their customers, from individual items to complex integrated systems. Their overall workflow breaks into four sequential major "phases." The first phase is initial interest from a customer in a broad product category, where the customer works with Broker to develop a structured expression of interest for a particular product set with a rough-order-of-magnitude (ROM) cost estimate. The second phase involves contract development (CD), where Broker develops and structures the contract to be executed in phase three, contract execution (CE). The fourth phase is contract closure. Elements of phase two and three are illustrated in the case study. Other phases are not shown for simplicity.

We provide two examples of workflow redesign in a case study. The first example introduces intra-phase concurrency in phase two, illustrated notionally in Fig 8, with a goal to reduce processing time within phase two. The second example introduces concurrency in inter-phase processes to convert certain non-deterministic processes to deterministic to reduce later contractual rework that causes multi-month delays. The goal in inter-phase workflow redesign is to reduce total system time.

A. *Intra-phase concurrency*

Phase two workflow translates the ROM developed in phase one into a formal itemized breakdown of costs and other contractual terms. Some phase 2 actions involve communications with US suppliers to get accurate cost and delivery information, interactions with US regulatory agencies to ensure appropriate legal authority to export is maintained, and interactions with customers. All new orders, whether for component parts or complex systems, flow through phase two.

In addition, modifications or amendments to existing orders also require the modified or amended orders to be reviewed through all phase two steps again to ensure the updated order is capable of being fulfilled during CE in phase three. Modifications to orders may be initiated in phase three for many reasons and may include, for example, transportation changes, corrections of misspecification of items or contractual terms, inadvertent errors, or changes in delivery time for one or more items in a particular order. The interactions with customers during CE quite often yield an amendment of the order that introduces a new workflow so the amended order can be properly reviewed and re-signed before once again proceeding through CE.

The duration of orders proceeding through phase two and three may span anywhere from several months to two years or more, depending on whether a particular order involves spare components parts, one or more major systems that are built to order, or a complex system comprised of complex machinery, on-site training upon delivery, and subscriptions for ongoing upgrades to technical manuals. With this complexity and multi-year order duration, modifications and amendments to orders comprise more than half of all orders. Thus one order may proceed through process workflows in phases two and three multiple times over the contract duration.

Our task was to map and model workflows to estimate Broker's employee workload levels. A major difficulty for

Broker, particularly for complex integrated systems, is long lead-times associated with receiving 1) detailed cost and contract data from US suppliers, and 2) appropriate documentation for legal export from regulatory agencies. While the workflows are largely deterministic, the stochastic nature of time to receive responses from suppliers and customers means that the employees may be nearly idle while in the midst of one of these lead times, or extremely busy in transitioning from one customer action to processing an action for another customer. The latter may induce delays due to bottlenecks in flow dependent actions, which extends lead times and induces additional costs.

Broker's performance is evaluated by many customer-oriented metrics, including total processing time for each phase. Broker recently instituted targets for reducing total time for processing workflow in phase two for very complex
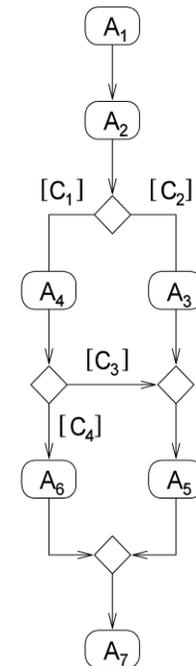
Fig.8.   A real-world case study workflow

systems from l_1 to l_2, a reduction of 25%, which is measured in months. If the flows of control of phase 2 can be reengineered from serial to concurrent, reductions in processing time will help reduce maximum lead times to l_2. As such, reengineered workflows would reduce total lead time by days or weeks, and thus 1) greatly improve responsiveness of Broker and Broker's suppliers to customers, 2) reduce task load imbalances on Broker's employees, and 3) reduce cost to Broker.

Fig 8 shows the workflow based on the case study. Due to the confidentiality agreement, we cannot disclose the actual steps and the inputs and outputs of the workflow. Instead, we replace them by generic names. The workflow in Fig 8 cannot be hastened because, except for the sequence A_1, A_2, there is only a single action node in between any pair of branch node or merge node. However, a more in-depth analysis reveals that A_3 is not atomic. In fact, A_3 can be decomposed into three pairwise

independent action nodes $A_{3.1}$, $A_{3.2}$, and $A_{3.3}$. As a result, they can be executed concurrently, as shown in Fig 9. Likewise, A_5 can be decomposed into three atomic action nodes $A_{5.1}$, $A_{5.2}$ and
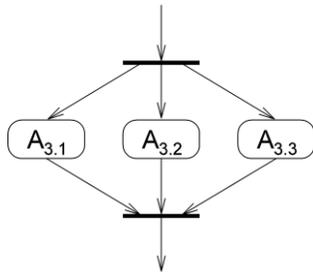
Fig.9. Rearranging three independent action nodes of A_3
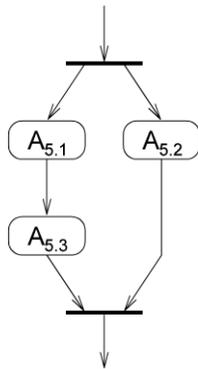
Fig.10. Rearranging three action nodes of A_5

$A_{5.3}$ of which $A_{5.1}$ and $A_{5.2}$ are not independent while the other pairs are independent. Thus, they can be rearranged into a workflow like the one in Fig 10.

The result of these reengineered workflows is reduced processing time to complete phase two, consistent with organizational goals, without an increase in resources or violations of workflow requirements or constraints. This example serves to illustrate how analysis and decomposition of processes provides opportunities to introduce workflow concurrency to serial tasks.

Currently, the principal barrier to introducing these concurrencies is lack of authority on the part of mid-level employees to move forward without prior approval by higher-level employees. For example, by policy $A_{3.2}$ currently has to wait on high-level approval on $A_{3.1}$ before it may proceed, although there is no strict reason, from the point of view of available resources $r_1 \dots r_n$ other than the approval itself (r_j). We have therefore recommend Broker review some of the policies that introduce serial approval steps to evaluate their necessity.

Our analysis also identified opportunities where inter-phase workflows across phases two and three are designed to be independent, but analysis revealed that a significant percentage of contract modifications and amendments are the result of

internal changes in how contracts are executed, as well as misspecifications and errors identified earlier require modifications and amendments above and beyond those initiated by the customer. The resultant workflows increase resource requirements, increase total processing time, and consume capacity that is constrained during fluctuations in daily and weekly workflow. We analyze this case in the next example.

B. *Inter-phase concurrency*

In addition to introducing concurrency to improve processing time in serial processes, such as in the phase two example above, our analysis identified non-serial processes in separate phases that unintentionally introduce non-standard practices in both phases. As noted, this causes inconsistencies and errors that make specific workflow processes non-deterministic. Due to the complex nature and long duration of many of these contracts, it is not surprising that these problems occur. But when they occur and contract modifications or amendments are necessary, the updated version of the contract must be re-processed through all of the phase two processes before the updated contract can be executed. This causes another full pass through phase two and through many of the processes in phase three. It is estimated the added workload is in the range of thirty percent of total workflows. In addition, often months can pass while the contract is re-specified, reviewed and re-signed. This affects resource costs, system capacity, total processing time, and customer satisfaction.

To address these problems, the goal in this workflow redesign example is to make more of the phase three processes deterministic by using workflow redesign to leverage the benefits of concurrency in phase two, specifically in contract review, verification, and structuring, to eliminate many of the latent contract problems identified only during CE in phase three. The original workflows are show in Fig.11 where the relevant processes for phase two are denoted as $A_7 - A_{12}$, and the processes for phase three are shown as $B_1 - B_8$. The workflows for phase two and three are f_5 and f_6, respectively. The final phase, contract closure, is not addresses in the process redesign and is denoted C_1 for simplicity.

The workflows of redesign interest in Fig.11 are discussed next. A_8 denotes a review process of customer requirements for a particular order by a CD manager ("CDM"). Due to the decreased phase two processing time targets for process time imposed by senior management, CDMs perform these review with input from suppliers that involve particular supplier requirements prior to forwarding the finalized customer requirements to the next phase. In A_{10}, the CDM translates these requirements from A_8 in to detailed contract requirements. In A_{11}, contract requirements are forwarded to the next process where the contract requirements are embedded in a finalized contract, ready for customer signature.

These three processes in the latter part of phase two are targeted for workflow redesign because many of the latent errors and misspecifications that arise months later in CE during phase three could be eliminated or mitigated during phase two.

It is noted that when working with organizational processes rather than information processing technology, it can be difficult to convert processes to deterministic. This challenge increases when a team that manages one phase for a key customer in phase two are located physically apart from a related customer team that manages in phase three. This example in our case study investigates how information uncertainty and inconsistencies between related contract specification teams and corresponding contract execution teams cause delays in contract execution,

We employ workflow concurrency as follows to accomplish this objective. For the workflow redesign of processes A_8, A_{10} and A_{11}, we first decompose workflow f_5 in Fig 11 into separate workflows and label them f_{5A}, f_{5B}, f_{5C}, f_{5D} and f_{5E} and f_{5F} (Fig 12). These workflows correspond to the six phase two processes identified in Fig 11 namely A_7, A_8, A_9, A_{10}, A_{11} and A_{12}, respectively. The three workflows, f_{5B}, f_{5D} and f_{5E}, will be redesigned in this example.

For the first workflow redesign in process A_8 (Fig 12), we introduce concurrent workflows using a fork node following A_7, where the new concurrent workflows denoted $f_{5B.1}$ and $f_{5B.2}$ correspond to processes A_8 and B_{R1-8}. This preserves the phase two review in A_8 by the CDM and introduces a concurrent review, denote B_{R1-8}, by the CE manager ("CEM"). The CEM uses a standardized review process that scans and identifies common errors, misspecifications, and other contractual problems that later drive increased workload.

The CEM review in B_{R1-8} in Fig 12 includes issues typically identified in phase three during any of the processes B_1 through B_8 that could have been avoided through better initial specification of customer requirements and how they match Broker's organizational capabilities to meet the requirements. These review are processed through a join node into a final customer requirements document (not shown for simplicity).
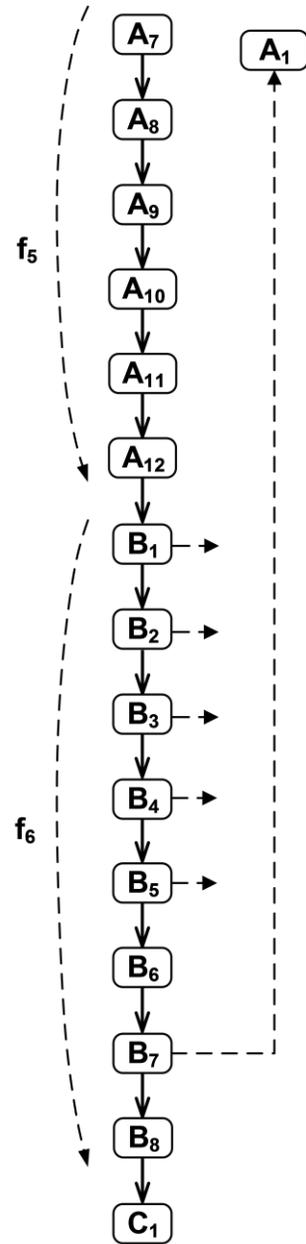
Fig.11. Original workflows for phase two (f_5) and phase three (f_6).
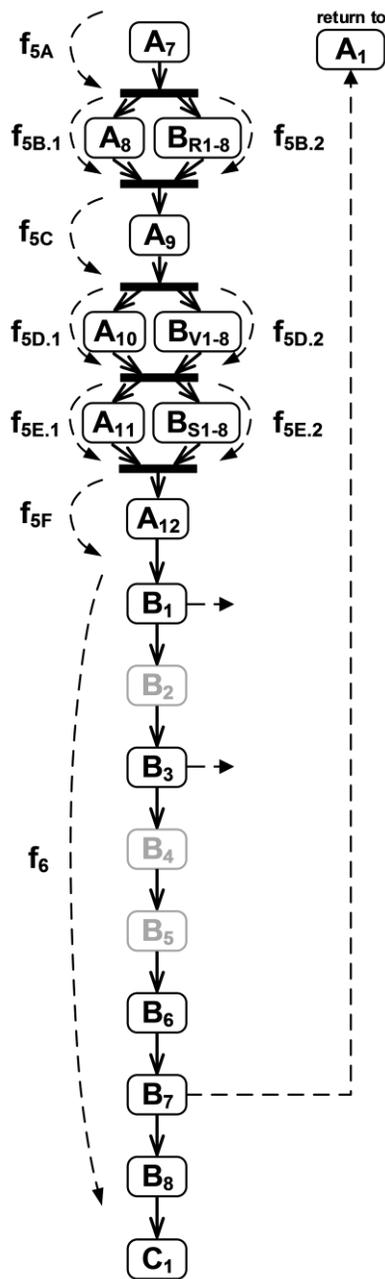
Fig.12. Redesigned intra-phase workflows using concurrency.

The second workflow redesign after the fork node following A_9, denoted $f_{5D.1}$ and $f_{5D.2}$ in Fig 12, introduces a concurrent review to verify detailed contract requirements in A_{10} and B_{V1-8}, which is driven by customer requirements derived in process A_8. Thus process A_{10} is also preserved, and the concurrent CEM review, B_{V1-8}, involves a standardized review to identify problematic contractual clauses and misspecifications in the draft of detailed contract clauses to ensure the contract clauses conform to customer requirements, Broker and supplier capabilities, relevant regulatory requirements, and international trade documentation and transportation requirements.

The detailed contract requirements resulting from A_{10} and B_{V1-8} in Fig 12 join after these processes, and immediately fork again to incorporate a concurrent review of the contract structure prior to final signatures. The CDM continues to perform a review of the contract structure in A_{11} while the CEM concurrently performs a review of known issues in contract structures that can affect phase three processing time, resource requirements, and customer satisfaction. Process flow then proceeds to A_{12}.

These concurrent reviews introduced in phase two are considered to be largely resource neutral, in that for most major contracts they already occur as part of a phase three process to identify problems that will occur later in time. Not all of these are caught in the phase three reviews during CE until they are ready for execution. In the longer term, concurrent reviews may be resource reducing when problems caught early during contract development eliminate the later need to expend resources and time to correct the contract or related supplier issues. Examples of these are shown in Fig 12, where processes B_2, B_4 and B_5 are represented in grey, signifying that these later processes do not need to be invoked for a particular contract if latent contract problems were identified and corrected in the concurrent reviews in phase 2.

C. *Benefits of concurrent reviews*

Introducing concurrent reviews for specifying customer requirements, detailed contract clauses, and contract structure much earlier in the process accomplishes several objectives. First, it identifies problems earlier, allowing more time to effectively react at lower cost while trying to maintain high levels of customer satisfaction. Second, by identifying and rectifying problems prior to formal contract signing, it avoids introducing contract modifications and amendments that must be processed through the full processes of phase two. This considerably reduces total process time to perform on the contract and initiate contract closure, reduces cost and resource consumption, reduces over-capacity utilization through reduced work flows, and improves customer satisfaction and Broker morale through improved customer performance and reduced execution of wasteful practices. It also avoids the requirement to run all contract modifications and amendments through all phase two processes, regardless of the extent of the contract modification or amendment.

These examples serve to introduce two type of concurrency that can be employed in organizational processes, with an intra-phase example of concurrency that employs decomposing serial processes into concurrent processing, and an example of non-serial inter-phase concurrency being employed earlier in organizational processes to convert certain non-deterministic work flows into deterministic workflows.

It is interesting to note that these workflow redesign opportunities existed prior to Broker introducing the phase two tightening of process performance times with the intent to better serve the customer. Thus the new shorter lead times to contract signature and execution did not create these problems, but they did exacerbate them. If concurrent workflows are introduced at processes A_8, A_{10} and A_{11} in phase two and maintained to effectively resolve the issues identified in this paper, the additional time to process these concurrent reviews

is on the order of days or perhaps a week or two, while the delays that the latent problems induce can be measured on the order of six to twelve months in additional delays. However, because of the current physical, organizational and social separation between the CDMs and CEMs, Broker is facing resistance in increasing the role of the CEMs in phase two.

This illustrates that the most effective policy to improve customer satisfaction is not to optimize time in a single phase at the expense of subsequent phases, but to consider the whole system and work to reduce total time from contract development to contract closure. Using concurrency in workflows can be effective in real organizations, and non-deterministic work flows can, in some cases, be converted to deterministic processes to reduce processing time and resource consumption while simultaneously improving customer satisfaction.

V. CONCLUSIONS & FUTURE RESEARCH

Adopting some of the concepts of activity diagrams, this paper defined a mathematical foundation for workflows. It then defined deterministic workflows, a class of workflows that have the predictable property that they will produce the same results if they have to be executed over and over again. After which, it defined flow independent workflows, a class of workflows that are deterministic. A methodology was then presented, which can be applied to each flow of control of a flow independent workflow. The methodology rearranges the action nodes in a flow of control so that some action nodes can be executed concurrently. We also supplement the paper by a real-world workflow that demonstrates the usefulness of our methodology.

Although introducing concurrency to flow independent workflows can be done in polynomial time, ultimately the workflows must be executed by humans or machines or any combination of the two. Hence, another relevant research problem is to assign the action nodes of a flow independent workflow to processors, which will eventually be responsible for executing the action nodes.

However, the problem of assigning action nodes of a flow independent workflow to a fixed number of processors so that the completion time of the workflow is under a certain time constraint does not seem to have a polynomial time solution. In fact, it seems like the problem is NP-complete. One can easily see the problem that given a number m of action nodes of a flow independent workflow that are pairwise independent, assigning them to $n < m$ processors to be executed so that the completion time of the workflow is under a specific time constraint subsumes the bin-packing problem, a well-known NP-hard problem. (In the bin packing problem, objects of different volumes must be packed into a finite number of bins or containers each of volume v in a way that minimizes the number of bins used.) We will report the investigation of this problem in a future journal paper.
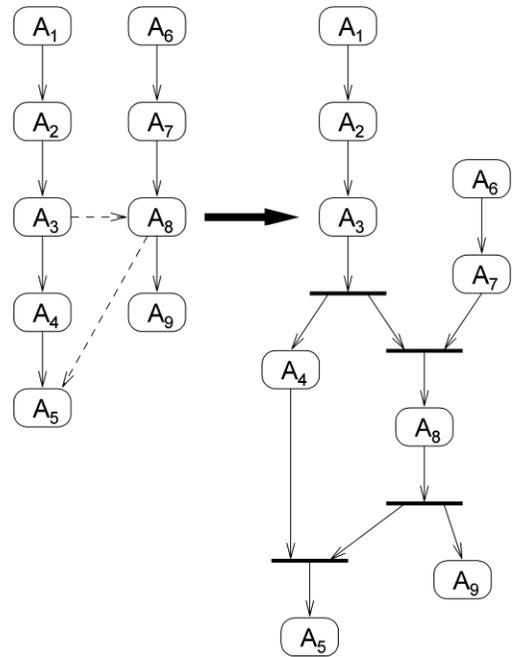
Fig.13. Adding fork nodes and join nodes to a not necessarily flow independent workflow.

Although flow independent workflows have the desirable property that being deterministic, as illustrated in this paper not all real-world workflows are flow independent. Another type of workflows have collaborating concurrent flows of control that pass data back and forth with one another. Although they may not be deterministic, nevertheless collaborating flows of control are quite common. Therefore, it is still useful to apply our methodology to a not necessarily flow independent workflow to speed it up.

The first step, however, is to identify the parts of the workflow to which our methodology can be applied. Although much more work is needed, we do have some preliminary ideas on this third approach. Fig 13 shows a workflow with two flows of control that are not independent. The dashed lines in Fig 13 denote data are passed from a sender action node to a recipient action node, and the recipient action node must wait for the data become available from the sender. In Fig 13, A_8 requires some data items from A_3 and A_8 send some data items to A_5.

Fig 13 also shows that after a careful analysis of the situation, in fact we can add fork nodes and join nodes to make the explicit modeling of passing data unnecessary. In addition, the added fork nodes and join nodes also delineate the sequences of action nodes to which our methodology can be applied. As shown in Fig 13, our methodology can now be applied to the sequences A_1, A_2, A_3 and A_6, A_7. We will also report these efforts in a future journal publication.

REFERENCES

[1] I. Arpinar, U. Halici, S. Arpinar, and A. Dogac, "Formalization of workflows and correctness issues in the presence of concurrency," vol. 7, pp. 199-248, April, 1999.

[2] P. A. Bernstein, and N. Goodman, "Concurrency Control in Distributed Database Systems," vol. 13, pp. 185-221, June 1981.

[3] G. Booch, J. Rumbaugh, and I. Jacobson, The Unified Modeling Language User Guide, 2nd edition, Addison-Wesley, 2005.

[4] F. Cédric, F. Dirk, and V. Hagen, "The relationship between workflow graphs and free-choice workflow nets," Information Systems, in press.

[5] Y. Gao, H. Ma, H. Zhang, X. Kong, and W. Wei, "Concurrency Optimized Task Scheduling for Workflows in Cloud," 2013 IEEE Sixth International Conference on Cloud Computing, pp. 709-716.

[6] M. Hammer, "Reengineering Work: Don't Automate, Obliterate," Harvard Business Review, vol. 68, pp. 104-112, Jul/Aug1990.

[7] J. Puustjarvi, "Workflow concurrency control," Computer Journal, vol. 44, pp. 42-53, 2001.

[8] X. S. Sherry, and J. L. Zhao, "Formal Workflow Design Analytics Using Data Flow Modeling," Decision Support Systems, vol. 55, pp. 270-283, Feb 2014.

[9] Y. Wang, and P. Lu, "Dataflow detection and applications to workflow scheduling," Concurrency And Computation-Practice & Experience, vol. 23, pp. 1261-1283, Aug 2010.

[10] Y. Wang, and P. Lu, "DDS: A deadlock detection-based scheduling algorithm for workflow computations in HPC systems with storage constraints," Parallel Computing, vol. 39, pp. 291-305, 2013.

[11] M. Wang, X. Zhang, L. Zhu, and L. Liao, "Trust-based workflow refactoring for concurrent scheduling in service-oriented environment," Concurrency And Computation-Practice & Experience, vol. 25, pp. 1879-1893, 2013.

[12] Object Management Group, "The Unified Modeling Language™ - UML," http://www.uml.org/

Quantum Gravity Sensor by Curvature Energy: their Encoding and Computational Models*

Francisco Bulnes, PhD, PostDoct

Head of Research Department in Mathematics and Engineering
Technological Institute of High Studies of Chalco
Chalco, P. C. 56641, State of Mexico, Mexico

Abstract—**Through of the concept of curvature energy encoded in non-harmonic signals due to the effect that characterizes the curvature as a deformation of field in the corresponding resonance space (and an obstruction to the displacement to the corresponding shape operator) is developed and designed a sensor of quantum gravity considering the quantized version of curvature as observable of gravitational field where the space is distorted by the strong interactions between particles, interpreting their observable in this case, as light fields deformations obtained on space-time background. To the application of this measurement we use a hypothetical particle graviton modeled as a magnetic *dilaton* which must be gauge graviton (gauge boson). Also are obtained several computational models of these photonic measurements, likewise their prototype photonic devices.**

Keywords—Curvature energy; magnetic dilaton; quantum gravity sensor; strong interactions; quantum computing

I. INTRODUCTION

The gravitation is the study of engendered field for the presence of matter in the Universe. This study begins with the determination of a field G, , and their sections $G_{ij} (i, j = 1,2,3,\ldots,N)$, , in the space through the universal gravitation law applied globally and locally, this last, considering the earthy ambit where is computed the acceleration in free fall of all bodies that fall in the event horizon of the Earth.

Subsequently and when the inertial effects due to the field and considering their coordinates system, likewise their transformations are established the relativistic models that consider chart systems (inertial reference systems of Galileo Galilei) to the determination of the field and their effects in said reference frame.

The study initiated with the universal gravitation is re-taken by Einstein in the frame of the curved geometry defining a geometrical stage for the only presence of the gravitational field G, , defined this stage for the curvature, taking the Einstein field equations to symmetric metric tensor :

$$R^{\alpha\beta} - \frac{1}{2} R g^{\alpha\beta} = -\chi T^{\alpha\beta}, \qquad (1)$$

being $R^{\alpha\beta}$, the curvature tensor or Riemannian tensor of the space-time. The constant χ, is the gravitation constant given by Einstein equal to $8\pi G / c^2$.

We can simplify the field equations as

$$R + \chi T = 0, \qquad (2)$$

which establishes that the space-time experiments *a curvature in presence of energy*. That is to say

$$R = -\chi T, \qquad (3)$$

The negative sign in $-\chi T$, is a convention, but can serve to remember us that the energy curves to the space-time to inside.

Then as is had said the gravity "winging" to the things to their source. A little change in the paradigm here is that the mass not is attracted between these "wing" on the space-time. These, curve to the space-time.

One important observation is that when is found with χT, must be imagined a spherical density. In change, when we find with R, the curvature is of the space-time as deformation of the spatial stage due the presence of matter. The exact values not are important in this last description.

In the non-Euclidean space-time or Riemannian space the symmetric tensor of the metric is $o^{\mu\nu(S)}$. This is the same that " g " the metric tensor but referred to distances where distance is symmetric as functional. Here the super fix " S " means symmetric.

The asymmetric tensor of the more general metric (*which could consider electromagnetic fields as field gauges to measurements of field*) is defined through the external or tensor product between tetrads (four metric vectors), that is to say:

$$o_{\mu\nu}^{ab} = o_\mu^a o_\nu^b = o_{\mu\nu}^{ab\,(S)} + o_{\mu\nu}^{ab\,(A)}, \qquad (3)$$

Asymmetric form haves inside herself as well as symmetry and anti-symmetry. The anti-symmetric component from (3) is $o_{\mu\nu}^{ab\,(A)}$. The anti-symmetric tensor of the metric is defined through the wedge product of two tetrads:

$$o_{\mu\nu}^{ab\,(A)} = o_\mu^a \wedge o_\nu^b, \qquad (4)$$

The action of the product of the tensors of curvature $R^{\alpha\beta}$ and o_μ^α, will establish an action of "torsing the action of gravity", which can be measurable *as distortions produced from the gravity.*

Then with the appearing of the quantum mechanics, more specifically, the QFT, and their interrelation with the gravitation is searched establish the cause of the field through the quantum interactions that generate this.

Then in this new "exhibition of gravity" the Einstein field equations (1) can be re-written as:

$$R_\mu^a - \frac{1}{2} R o_\mu^a = \chi T_\mu^a, \qquad (5)$$

and using the fact that now our new metric tensor can be expressed as[1]:

$$g_{\mu\nu}^{ab(A)} = o_\mu^a \wedge o_\nu^b, \qquad (6)$$

we arrive to the new field equation to electrodynamics that is generally covariant:

$$o_\mu^a \wedge (R_\nu^b - \frac{1}{2} R o_\nu^b) = \chi o_\mu^a \wedge T_\nu^b, \qquad (7)$$

which give us the spin or torsion of the field.

In this asymmetric space-time model are obtained field models that reflect this torsion. For example the Asymmetric field theory given by Yang-Mills where this theory provides an extension of Maxwell theory to the case of non-Abelian fields. In this dimensions raise the wrappings and the loop contributions that will contribute to the energy micro-states used to define electromagnetic signal effects of power that can consigned in a harmonic analyzer.

Subsequently with the string theory [1] and wanting to solve the problem of the super-symmetry, is developed the concept of quantum gravity which establish through of hypothetical particle called "graviton" [2] (which by the duality principle of the QFT, on field/particle) could establish the origin of the field $G,$ as the bosonic source of the gravitation.

After and under studies of the QED, are established indirect methods to the determination and detection of quantum gravity using gauge fields or gauge bosons respect to the background radiation of the Universe and these produce for "backreaction" a measure of the distortion that produce the presence of the field $G,$ as the causing of the variation and distortion of the gauges field with the microwaves of the space-time.

Then these variations can be the pattern that the Cartan tetrads $o_{\mu\nu}^{ab},$ define to realize the measurements of quantum gravity fluctuations and that can be the solutions to one

obtained quantum field equation, analogous to the Dirac equation considering the electromagnetic and gravitational fields [2] together and whose solutions can be measurable gravitational waves in electromagnetic waves through of filters of signals:

$$(\Box + \chi T) o_\mu^a = 0, \qquad (8)$$

II. H-Fields in a Generalized Curvature Tensor and First Boson-Fermion Measurements

Considering the integration formalism through the total action integral of $g_{\mu\nu}^{ab(A)}$, that comes from the sum of the partial actions due to the curvature tensor and of the electromagnetic tensor include in this last the fermion self interactions induced by quantum torsion we have that a second integral of (7) is:

$$\Im_{TOTAL} = \frac{1}{2\kappa} \int d^4 x o o_\mu^a o_\nu^b R(\varpi) + \frac{i}{2} \int d^4 x o \times$$
$$\times (o_\mu^a (\overline{\psi}\gamma^a \mathcal{D}_\mu(\varpi, A)\psi - \overline{\mathcal{D}_\mu(\varpi, A)\psi}\gamma^a \psi)) -$$
$$- \int d^4 x o \frac{3}{10} \kappa J_{(A)}^\mu J_{(A)\mu}, \qquad (9)$$

Finally this global action defined in (9) can be re-written to fermions in gravity with torsion [3, 4], with a specific kind of torsion (*Kalb-Ramond field strength* [5]) inspirited from the string theory mentioned before (UV complete) can do the job of providing a constant , axial background in a local frame of FRW − cosmology. Then the additional fermion-piece of the form

$$\Im = \frac{\alpha}{2} \int d^4 x o (\overline{\psi}\gamma^a \mathcal{D}_\mu(\varpi, A)\psi - \overline{\mathcal{D}_\mu(\varpi, A)\psi}\gamma^a \psi),$$
$$\alpha = cte, \qquad (10)$$

together with Dirac kinetic terms , the fermion action reads:

$$\Im_{\text{Dirac-Holst-Fermi}} = \frac{\alpha}{2} \int d^4 x o (\overline{\psi}\gamma^a (1 - i\alpha\gamma_5)\mathcal{D}_\mu\psi -$$
$$\overline{\mathcal{D}_\mu\psi}\gamma^a (1 - i\alpha\gamma_5)\psi), \qquad (11)$$

where inside the integrand are involved the Dirac equations to the differentiating fermions in the non-Harmonic analysis

---

[2] Gravity with torsión contains anti-symmetric parts in the connection:

$$\omega_\mu^{ab} = \varpi_\mu^{ab} + K_\mu^{ab},$$

where $\varpi_\mu^{ab} = o_\mu^a \partial_\mu o^{\nu b} + o_\nu^a o^{\sigma b} \Gamma_{\sigma\mu}^\nu = o_\nu^a o_{;\mu}^{\nu b}$. Then we have a co-torsion tensor $K_\mu^{ab} = o_\nu^a o_\rho^b K_{\nu\rho}^{ab}$, such that $K_\mu^{\nu\rho} = -K_\mu^{\rho\nu}$.

---

[1] That is to say, the new metric tensor is anti-symmetric.

that appear in the anti-symmetric behavior of the curvature field measured for quantum interactions (see Figure 1).
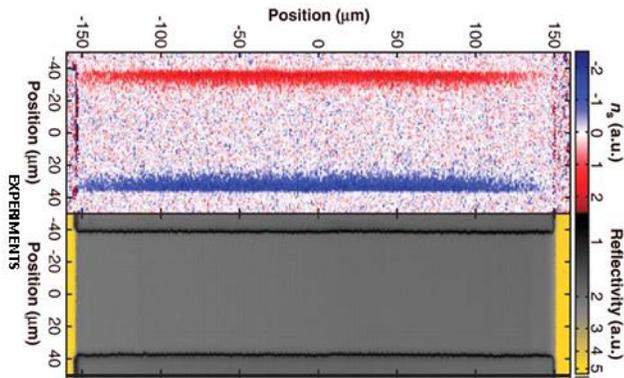
Fig.1. Differentiating fermions in the non-Harmonic analysis that appear in the anti-symmetric behavior of the curvature field measured for quantum interactions.

Inside of the microscopic UV complete underlying theory of quantum gravity if we consider the massless gravitational multiplet of *"closed" strings* such as spin 0, scalar or *dilaton*, spin 2, traceless symmetric rank 2 tensor or the *graviton*, and spin 1, antisymmetric rank 2 tensor or *electromagnetic tensor*, with the Kalb-Ramond field $B_{\mu\nu} = -B_{\nu\mu}$, we can have a gauge invariant to effective field theories (in low energy scale $E << M_s$) given for $B_{\mu\nu} \rightarrow B_{\mu\nu} + \partial_{[\mu}\theta(x)_{\nu]}$, which depend only on the field strength $H_{\mu\nu\rho} = \partial_{[\mu}B_{\nu\rho]}$. Then we give the Bianchi identity:

$$\partial_{[\sigma}H_{\mu\nu\rho]} = 0, \qquad (12)$$

But to the detected anomalies through the gravitational versus gauge interactions cancellations of strings (*necessaries to the apperceiving of the gravitational waves letting only the gravitational strings*) is required a re-definition of the $H-$ fields given in (12) using the extension due to the Majorana neutrinos masses from (three loop) anomalous terms with axion-neutrino couplings. Then the corresponding extended Bianchi identity to these anomalous terms come given as:

$$H_{\mu\nu\rho} = \partial_{[\mu}B_{\nu\rho]} + \frac{\alpha'}{2\kappa}(\Omega_L - \Omega_V), \qquad (13)$$

where result interesting study the phase-space density derived from the difference between of the Chern-Simons three forms [6] Ω_L, and Ω_V, where as Lorentz Chern-Simons three form Ω_L, is considered the corresponding to neutrinos, and to the case of the gauge Chern-Simons three form is considered the corresponding electro-gravitational formalism in gauge theory.
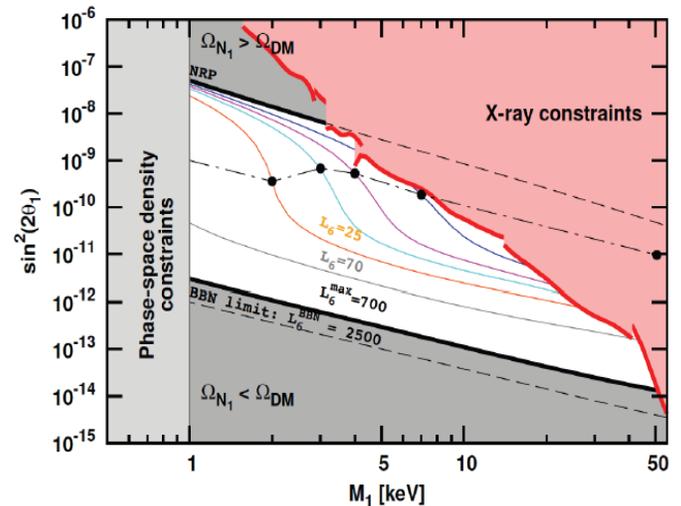
Fig.2. Leptogenesis/Baryogenesis model consistent with BBN- structure formation data in the Universe and all other astrophysical constraints [7].

A theoretical study related to the propagation of photons in quantum gravity, shows that the region of space- time of the mechanical well of a singularity is supported by an energy that decreases asymptotically in the infinite. This hypothetical energy we can construct it with the expression of a Lagrangian like the given in (Table 2 [8-10]) with cylindrical gravitational wave by the *dilaton* (gauge particle):

$$\Phi = (1/10000(\exp(-4\xi)J_{\nu,x}(3\xi,1) + \exp(-4\xi)Y_{\nu,x}(2\xi,1)), (12)$$

wave model for energy of gravitational waves (see last line of the Table 3)) [8]. Also see the figure 3 [8].
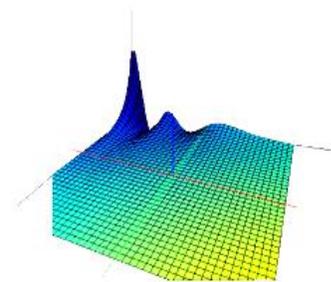
Fig.3. Gravitational alteration perceive by the censor designed by (13) when is obtained a great alteration of energy near the singularity of the space-time.

III. CURVATURE ENERGY TO DESIGN A QUANTUM GRAVITY SENSOR

According to our curvature studies that come from a theoretical sensor of curvature in presence of the incurve region of the space and detected by a light wave [8, 11], the curvature follows being the more important geometrical invariant that appear as a geometrical effect in the space when appears a variation of the field due to the deformations of the light geodesics encoding as light waves that transit near of material sources of space where

The data of these encodings where have been consigned in a first model of "gravitational waves" suggest that the detection

of gravity can be through of the energy obtained by a *cosmic censorship*, given by

$$[log\,\phi(\xi(s))]^2 \left[\int log\,\sigma(s)\right]^2 \geq (\int \Omega(1-\nabla^2 log\Omega))^2 \geq 4\pi\int\Omega,$$

(13)

This is a Hilbert inequality that through "curvature energy" measured with models of Gaussian and normal curvatures establishes an energy range where the curvature energy can exist if their Lagrangian haves a Maxwellian component. Observe the certain bound of roundness obtained through of the implementation of a spherical operator [9, 12], we can use one particular case given by measurement for light waves contemplated in the *Bulnes's theorem* published in *ASME* 2009, [11].

The possible integral expression (13) of the total Hamiltonian of electromagnetic energy establishes a condition of domineering energy [11], when there exists curvature in the 3-dimensional space (that is to say, if the energy is given by this *roundness censor*, there is curvature measured like energy that makes the censoring appear and measured directly using positioned of the sensor device). This makes possible to obtain a curve of perception of curvature that measures the condition of domineering energy in the sensor to establish their position in the surface [12] and of this way to measure in every point of the surface their curvature.

But the cosmic censorship, which due to their design through classical electromagnetic fields, requires of a super-massive source to obtain a significant variation of the gravity presence (see the figure 4).

The gravitational waves can be modeled by non-harmonic signals produced by detection of the super-massive source like a black hole in the Universe [8].
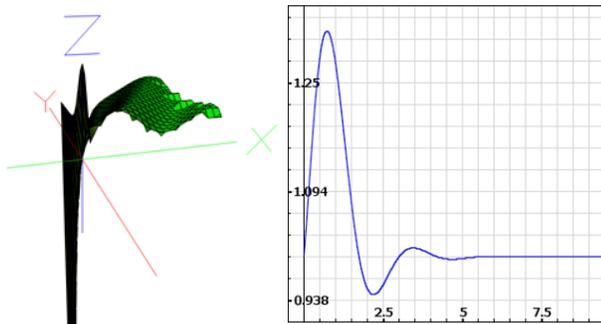
Fig.4.

Then using the difference of energy that can detect a sensor of gravity that re-collect little variations or fluctuations due to inertial effects of the matter existence in the space-time, we can measure curvature we base in the idea of consignee in angles from the resonances [11]. Such is the case, for example, the use of an inertial device as an accelerometer which with a capacity of dynamic perception given by their electrostatic force $F_e = 1/2\,AV^2/d^2$, whose deflection is consigned in a $g-$ cell component.
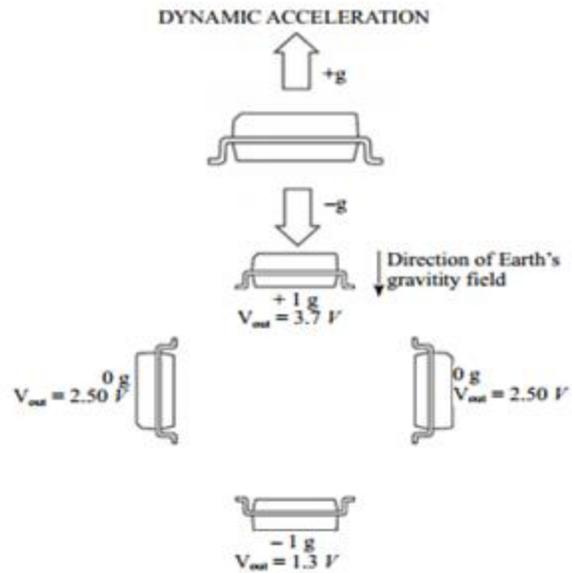
Fig.5. Accelerometer type model used to perceive variations of gravity in the usual sense [12]. This will be modified to can receive the fine variations or fluctuations between boson gauges and microwaves in background. Their gravity element will substitute by non-Newtonian fluid which wills consignee the little variations of gravity from the values of each parasite capacitors.

This force is designed in our sensor by the basic isotopic component of Gaussian *factor* to lectures of curvature defined as $\alpha(-1)^2 \bullet 1 \bullet (4\pi^2)$, where $(-1)^2$, is the basic charge given in function of the milimetric potential V, the factor $A = 4\pi^2$, that there is in the surface of the sphere of radius the unit $S^2(1)$, α, is the degree of the spherical map used in the transduction of the physical model to measure, which comes like a factor of electromagnetic adjustment of the sensor on the curved surface and the factor 1, is the positive charge generated inside the sensor (see figure 5). This is equivalent to the product of the force for the square of the distance of separation between the plates sensors of the nucleus of the sensor, where one gathers the change of the load according to the surface (negative load), and other gathers the charge invested in the process of detection (positive charge). The resultant deflection is measured by the accelerometer's control ASIC and a proportional output voltage results [12, 13]. This procedure assures that both the mechanical ($g-$ cell) and electronic sections of the accelerometer are functioning [14].

How measure curvature of the space-time from the concept of quantum gravity interpreting their observable in this case, as light field deformations obtained on space-time background?

Extra-poling the idea of the cosmic censorship to the quantum field fluctuations due to interacting of a *dilaton* with the microwave background we can detect through the difference scattering of fermions detected in each case; particles/anti-particles. In this case are neutrinos/anti-neutrinos where the curvature will be given for the sufficient quantity of sterile neutrinos that must appear to reproduce observed oscillations which express induced torsion. The apparent

flatness in the space is conserved and only in macroscopic region of the space-time is
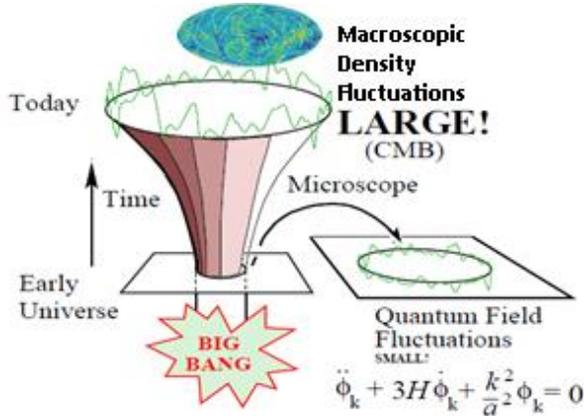
Fig.6.
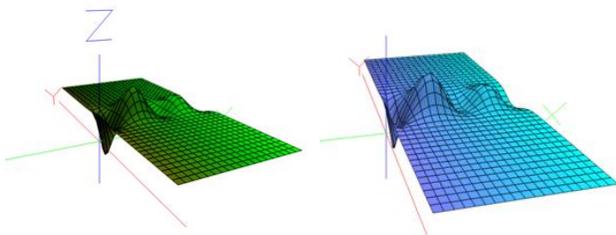
Fig.7. Modeling and measurements obtained by a theoretical sensor of quantum gravity designed accords with the quantum field fluctuations in the flatness of the space-time supported by the neutrinos/anti-neutrinos totality.

Using the concept of *curvature energy* mentioned in our introduction we can annunciate the following result which was published in [12]:

Theorem (F. Bulnes). III. 1. [12] Let $\mathfrak{a}$, be an accelerometer given by the device D, defined for their $g-$ cell whose curvature energy to the case $2-$dimensional is the bounded by the integral $\Lambda \oint_C k^2 ds$, [11, 15], where Λ, is an output electromagnetic factor from D, (to obtain curvature). Let $\mathbf{S}$, the corresponding shape operator on S^2 (and their contours), that is to say, is that whose normal curvature is given for $k(\mathbf{u}) = \mathbf{u}\mathbf{S}_p(\mathbf{u}) = \mathbf{u}(\text{-}1/r)\mathbf{u} = -1/r$. Then our curvature is defined for the inequality of voltage-curvature energy given by (1) whose energy is $\mathcal{E}_k = (1/2)V^2 A \sin^2 \theta$.

Proof. [12].

Difference of Energy can be consigned in the quantum distortion given by the link-wave between graviton modeled as dilaton (gauged graviton) and the trace on relativistic Feynman diagram followed in quantum gravity. The quantum curvature as a different times in the causality and conformally in the space-time. The different deviations to the world-lines in each case show the curvature: *spinor frame* [16, 17]

IV. PHOTONIC MEASUREMENTS AND COMPUTATIONAL MODELS

Newly considering the effective gravitational action in string low-energy in terms a generalized curvature Riemannian tensor, where the Christoffel connection includes the $H-$ fields, that is to say, $\Gamma_{\nu\rho}^{\mu} = \Gamma_{\nu\rho}^{\mu} + \frac{\kappa}{\sqrt{3}} H_{\nu\rho}^{\mu} \neq \Gamma_{\rho\nu}^{\mu}$, defined in (13), we can give the $4-$dimensional action:

$$S^{(4)} = \int d^4 x \sqrt{-g} \left(\frac{1}{2\kappa^2} R - \frac{1}{6} H_{\mu\nu\rho} H^{\mu\nu\rho} \right)$$
$$= \int d^4 x \sqrt{-g} \left(\frac{1}{2\kappa^2} \overline{R} - \frac{1}{3}\kappa^2 H_{\mu\nu\rho} H^{\mu\nu\rho} \right), \quad (14)$$

where the dual of H, in four dimensions comes given by the differential equation:

$$-3\sqrt{2}\partial_{\sigma}b = \sqrt{-g} \in_{\mu\nu\rho\sigma} H^{\mu\nu\rho}, \quad (15)$$

where $b(x)$, is a pseudo-scalar that defines the Kalb-Ramond axion. Then to a dilaton Φ, that satisfies (15) having the properties described in (8) as gravitational wave, we have the differential equation:

$$H^{\mu\nu\rho} = o^{2\Phi} \in_{\mu\nu\rho} \partial^{\sigma} b(x), \quad (15)$$

The linear dilaton solution in *string frame*, (or logarithmic in FRW time in Einstein frame) with conformally flat Einstein-frame target space-time is exact in all orders of a parameter , that appears in (11).

Using the principles dictated in (9)-(11) and the differentiating fermions in the non-Harmonic analysis that appear in the anti-symmetric behavior of the curvature field measured by quantum interactions we can give the following action that comes from of the Majorana states in fermionic field theories with $H-$torsion:

$$S_{\psi} = -\frac{3}{4}\int d^4 \sqrt{-g} S_{\mu} \overline{\psi}\gamma^{\mu}\gamma^5 \psi, \quad (16)$$

Using the extra-charge created by the fermion interaction (central charge of underlying world-sheet conformal field theory [16]) where

$$b(x) = \sqrt{2}e^{-\phi_0}\sqrt{Q^2}\frac{M_s}{\sqrt{n}}t, \forall n \in \mathbb{Z}, \quad (17)$$

and using said charge to create a basic charge in a component of $g-$cell, and also using our theorem on curvature considering an isotopic component of Gaussian *factor* to lectures of curvature we can define a sensor whose 3-ball of non-Newtonian fluid can receive these signal and re-interpreting through voltage-curvature energy, such and say the *theorem III. 1*, these data as little electrical tensions that come from the surface of the 3-ball which can be sensored as little

change in the background due to the dilaton in interaction with this.

An experiment is done with a little accelerometer that include a charged ball whose charge vary in the time when this accelerometer change their position respect to their horizontal frame (defined this for their Ecuador) given in their $g-$cell change (see the figure 8). We use two leds to establish the polarization effect created in natural form by the fermionic behavior.

The before can to help define a classical accelerometer in the earth's gravity. The curvature will be able to express itself like a Gaussian curvature according to spherical harmonics given by Legendre polynomials.

The sensor is a sensor of free fall that can register different force factors G. The difference is consigned by the Hall effect obtained by the difference scattering of fermions detected in each case; particles/anti-particles. The actions of change can be reprogrammed by the proper device considering these to be a Lagrangian action given for [8].
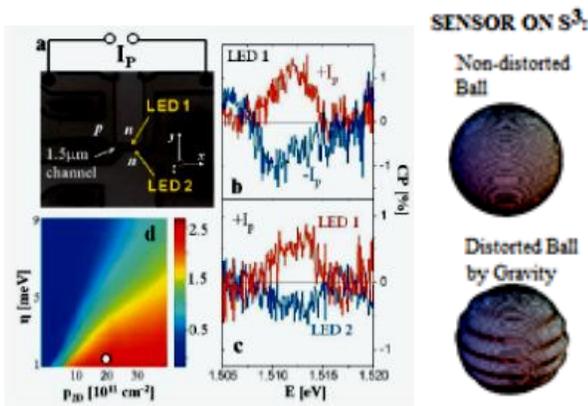
Fig.8. The difference is consigned by the Hall effect obtained by the difference scattering of fermions detected in each case; particles/anti-particles.

Then extrapolating this experiment in the ambit of the photonics the folds or "creases" in a deformable sphere are oscillations in the Universe which comes given by the mixture of neutrinos/antineutrinos for the eco of the Early Universe which maintained their basic non-spherical symmetry until our actual Universe and which can be expressed through of their Lagrangian as:

$$\mathcal{L} = \mathcal{L}_f + \mathcal{L}_I = \sqrt{-g}\,\overline{\psi}[(i\gamma^a\partial_a - m) + \gamma^a\gamma^5 B_a]\psi, (17)$$

where $\overline{\psi}$, and ψ, are component of the spinor Ψ. The oscillations are received as spherical auto-modes of the alteration of central charge Q, obtained by the differentiated fermionic process (see the figure 9) (extension of the model the axion $b(x)$ using total derivatives of the gravitational $cR^{\mu\nu\rho\sigma}\tilde{R}_{\mu\nu\rho\sigma}$, and electromagnetic $cF^{\mu\nu}\tilde{F}_{\mu\nu}$, [18] terms of the fields o_μ^a, translated to $H-$fields).
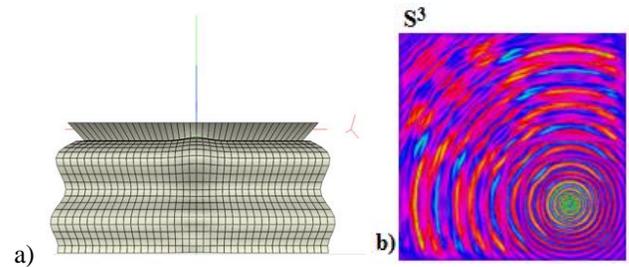
Fig.9. a). Comparison between the theoretical norm obtained by the connection given by (6) (*whose distance formula*) is (red curve) and WMAP-Satellite measurements (black and blue curve) b). Measuring the presence of cosmic microwave background (*CMB*) on the theoretical twistor surface of hypothetical ball S^3, in our sensor device. C). Image obtained by *WMAP* satellite [19].

The neutrinos/anti-neutrinos conform the asymmetry around black holes or singularities. Inside of singularities the gravitational field is dementia. Then their particle/anti-particle can be generated from the torsion. Using plane wave approximation is obtained different dispersion relations between particles/anti-particles at finite densities assuming constant background torsion (see the figure 10).
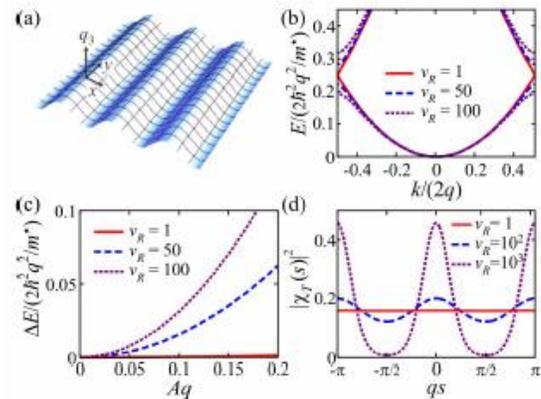
Fig.10. Quantized curvature energy in the harmonic case. a) The gravitational waves model is symmetric and obeys to the spherical symmetry. b), c) and d) Dispersion relations between particles/anti-particles.

Finally through a magnetic dilaton Φ, we can give a model of magnetic distortion, that is to say, the energy curvature in the gravitational media can be translated as magnetic deformation of $4-$dimensional part of the string of background radiation (see the figure 11).
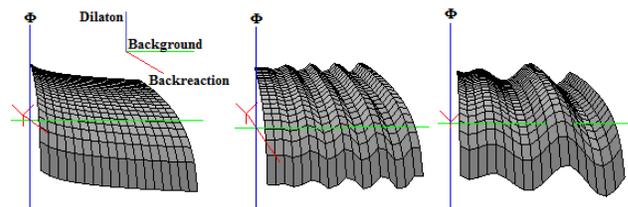
Fig.11. Dilaton measuring distortion due to quantum gravity, according to the model computacional magnetic $\Phi = y/\sqrt{y+1} - ((1/2)\log(x + 1/x + (x - 1/x))\cos\theta$ (see equation (19)). The surface in a), represents the space-time area before the photons back-reaction with background radiation, their magnetic model is $\Phi = y/\sqrt{y+1} - ((1/2)\log(x + 1/x + (x - 1/x)))$. In the surface b), the distortion is measured by the magnetic oscillations $\cos\theta$, that is to say, the term

of tough deviation θ, in the figure 1. In c), the distortion is attenuated by the background increasing the undulations (*cosmθ*, $\forall\ m \in \mathbb{z}^+$) and increasing their amplitudes being not detectable for being under the background (*green line*).

The gravitational energy is the curvature energy obtained through components Bessel functions or polynomials (see the figure 12).

A)

B)

$$y = 0.8 \cdot \exp(-0.05 \cdot t) \cdot \sin(t) + 0.8 \cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(3 \cdot t)}{3} + 0.8$$
$$\cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(5 \cdot t)}{5} + 0.8 \cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(7 \cdot t)}{7} + 0.8$$
$$\cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(9 \cdot t)}{9} + 0.8 \cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(11 \cdot t)}{11} + 0.8$$
$$\cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(13 \cdot t)}{13} + 0.8 \cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(15 \cdot t)}{15} + 0.8$$
$$\cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(17 \cdot t)}{17} + 0.8 \cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(19 \cdot t)}{19} + 0.8$$
$$\cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(21 \cdot t)}{21} + 0.8 \cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(23 \cdot t)}{23} + 0.8$$
$$\cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(25 \cdot t)}{25} + 0.8 \cdot \exp(-0.05 \cdot t) \cdot \frac{\sin(27 \cdot t)}{27}$$
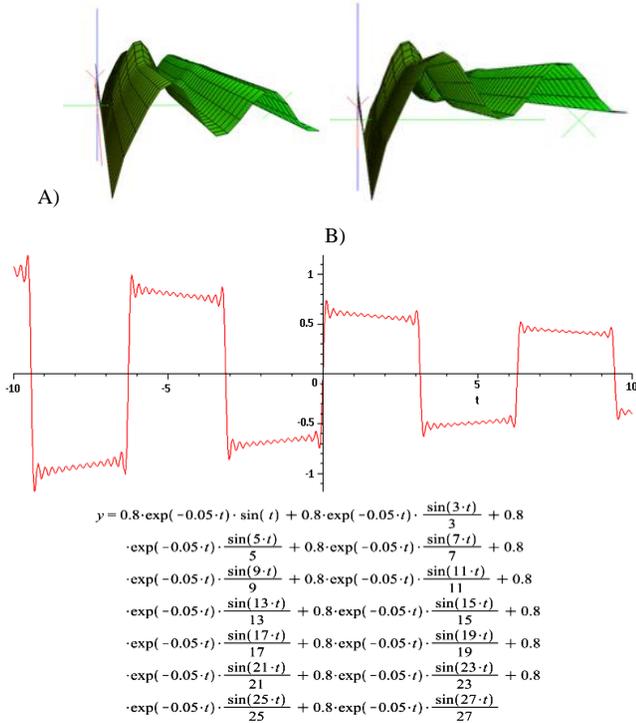
Fig.12. A) Curvature energy surfaces. The curvature represents an *obstacle* of the field flow and the "work" realized by the field to move their flow is translated in energy of curvature. B) Rescaling of the signal we can obtain a Bessel spectra corresponding to the energy curvature. The numerical operative sequences are realized to obtain this spectrum [20, 21].

V. CONCLUSIONS

The obtaining of geometrical models of quantum gravity comes to show the quantum behavior of observable obtained for photons that act on the background radiation (or microwave radiation) such as the torsion induced for H − fields that through dilatons or gauge bosons can exhibit observables of gravitational field as curvature to quantum gravity, where the curvature in this last case is detected and viewed as curvature energy or gravitational energy that exists until our days as a eco of the times of the creating of the gravity in the step of a Early Universe.

Then the observational confirmation due to the gravitational background generating CPT violating effects in the Early Universe define experimentally by the particles/anti-particles differences in dispersion which implies CPTV effects of different space-time-curvature/spin couplings to neutrinos/anti-neutrinos [22].
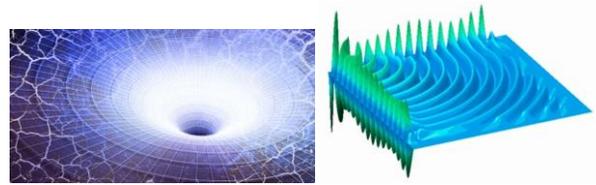
Fig.13. Gravitational alteration perceive by the quantum gravity sensor designed by the charge alteration in the fermion couplings. The waves that are *spinors* can be consigned in oscillations in the *space-time-curvature/spin*. The modeling of spinor "quantum waves" was done in ANSYS.

The curvature coupling to fermion spin (see the figure 14) may lead to different dispersion relations between neutrinos and anti-neutrinos (assumed dominant in the early eras): in non-spherically symmetric geometries in the Early Universe.
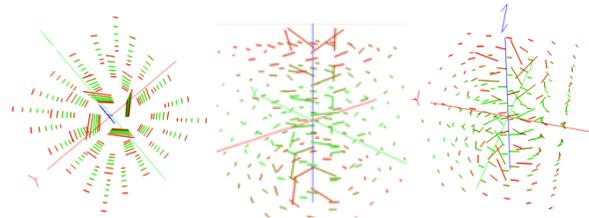
Fig.14. The induced torsion for non-spherically symmetric geometries by the curvature coupling to fermion spin. The attractor hole in the figure 13, can be constructed from the fermion couplings showed in these computational models.

Such geometries can be consigned sensing and perceiving through sensor whose g − cell include a hyper-sensible component that consists of a "sphere" composite of a non-Newtonian fluid (that perceive the Gaussian curvature units) which can be deformed by variations of gravity to start of the values of each parasite capacitance. The little alteration of their surface in non-harmonic "creases" produces an alteration of their charge [12, 23, 24] (see the figure 14).
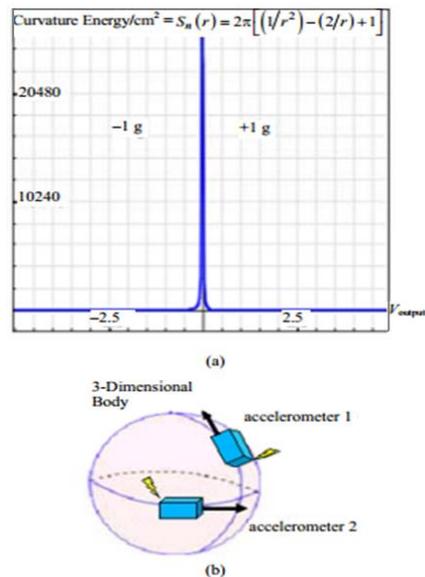
Fig.14. The curvature will be able to express itself like a Gaussian curvature according to spherical harmonics given by Legendre polynomials.

TECHNICAL NOTATION

K – Curvature as general concept of roundness property. Also used in the paper as Gaussian curvature in a point p

k – Gaussian curvature along of the geodesic or surface

$\mathcal{O}_E$ – Spherical operator which related with the spherical map.

k_i – Principal curvature in the principal ith-direction. In our research i = 1, 2, only

V – Voltage

A – Area

$\mathcal{L}(H(\Omega^2(M)))$ – Space of spectral transformations on curvature $2-$forms given in space $\Omega^2(M)$

M – Space whose curvature is measured. In our study M, represent $2-$dimensional surfaces or $3-$dimensional bodies.

$S^2 - 2-$Dimensinal sphere. Also is the $2-$dimensional sphere used in the spherical map to design our curvature sensor

$r(P)$ – Evaluation of curvature radius from the product from their inverse principal curvatures.

Ω – Curvature form in the 4, and $4-$dimensional spaces

$J_{1-r,p}$ – Bessel function of first specie

$Y_{1-r,1-r}$ – Bessel function of second specie

$SOIC$ – Small-outline integrated circuit

$EDMC$ – Electromagnetic device to measure curvature

mV – Micro-volts

(Z^α, W_α) – Ambitwistor element whose elements are invariant-covariant fields

$SU(2)$ – Group that defines the finite actions through unitary anti-Hermitians matrix of range 2

$ASIC$ – Application-specific integrated circuit

QED – Quantum Electrodynamics

QFT – Quantum field theory

$SUSY$ – Super-symmetry theory

FRW – cohomology- Friedman-Robertson-Walker metrics. The cosmological principle (principle of homogeneity and isotropy of the universe to great scale). The *cohomology* are the relations of similarity in dual spaces.

REFERENCES

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

[1] Blumenhagen, Ralph; Lüst, Dieter; Theisen, Stefan (2012), *Basic Concepts of String Theory*, Theoretical and Mathematical Physics, Springer, p. 487, "Orbifolds can be viewed as singular limits of smooth Calabi–Yau manifolds".

[2] T. Rothman, S. Boughn, (2006). "Can Gravitons be Detected?". *Foundations of Physics* **36** (12): 1801–1825. arXiv:gr-qc/0601043

[3] N. E. Mavromatos, S. Sarkar, "CPT-Violating Leptogenesis induced by Gravitational Defects," *High Energy Physics - Phenomenology* KCL-PH-TH/2012-23; LCTS/2012-28, pp1-22.

[4] J. Ellis, N. E. Mavromatos, S. Sarkar, "Environmental CPT Violation in an Expanding Universe in String Theory," *High Energy Physics – Phenomenology*, KCL-PH-TH/2013-13, LCTS/2013-07, CERN-PH-TH/2013-079, pp.1-8, arXiv:1304.5433

[5] M. Kalb, P. Ramond (1974). "Classical direct interstring action". *Phys. Rev. D* **9** (8): 2273–2284.

[6] S-S. Chern, J. Simons (1974), Characteristic forms and geometric invariants, The Annals of Mathematics. Second Series **99** (1): 48–69,

[7] A. Boyarsky, A. Neronov, O. Ruchayskiy, M. Shaposhnikov , "Restrictions on parameters of sterile neutrino dark matter from observations of galaxy clusters," Physical Review D, 74 (10) 103506.

[8] Bulnes, F. "Electromagnetic Gauges and Maxwell Lagrangians Applied to the Determination of Curvature in the Space-Time and their Applications," *Journal of Electromagnetic Analysis and Applications*, Vol. 4 No. 6, 2012, pp. 252-266. doi: 10.4236/jemaa.2012.46035.

[9] Van Vliet L. J.; Verbeek, P.W. Curvature and bending energy in digitized 2D and 3D images, in: SCIA'93, Proc. 8th Scandinavian Conference on Image Analysis, Tromso, Norway, 1993, 1403-1410.

[10] Król, J. Quantum gravity insights from smooth 4-geometries on trivial R⁴, Quantum Gravity, Intech. Chapter IV, pp53-78, Croatia, 2012.

[11] Bulnes, F.; Hernandez, E.; Maya, J. Design of Measurement and Detection Devices of Curvature Through of the Synergic Integral Operators of the Mechanics on Light Waves. Proceedings International Mechanics Engineering Conference and Exposition, Orlando Florida, USA, November 16; ASME, 2009, 91-102. http://dx.doi.org/10.1115/IMECE2009-10038.

[12] Bulnes, F. Martínez, I. Mendoza, A. and Landa, M. "Design and Development of an Electronic Sensor to Detect and Measure Curvature of Spaces Using Curvature Energy," *Journal of Sensor Technology*, Vol. 2 No. 3, 2012, pp. 116-126. doi: 10.4236/jst.2012.23017.

[13] Pallás, R. and Webster, J. G. *Sensors and Signal conditioning* 2nd ed. John Wiley & Sons, USA, 2001.

[14] Nashelky B. *Electronics: Circuit Theory and Electronic Devices,* 1st ed.; Pearson Prentice Hall, USA, 2003.

[15] Bulnes, F. *Research on Curvature of Homogeneous Spaces*, 1st ed.; TESCHA: State of Mexico, Mexico, 2010; pp. 44-66.

[16] A-Wollmann Kleinert and F. Bulnes, "Spinor Technology: the Field Formalism to the Duality between Quantum Particles and their Waves," Journal on Photonics and Spintronics, Vol. 2, n. 2, 2013, pp9-18.

[17] A. Wollman-Kleinert, F. Bulnes and D. Guderian, **"**Measurable Quantities of Quantum Field Theory: An Reviewing of Scalar Field Theory and their Applications," *Journal on Photonics and Spintronics* Vol. 3, no. 2, 2014, pp.10-19.

[18] Bulnes, F. Doctoral course of mathematical electrodynamics, Internal. Proceedings Appliedmath 2, SEPI-IPN and IM-UNAM, México City, Mexico, 2006, 398-447. ISBN: 970360358-0.

[19] Smoot Group, "The Cosmic Microwave Background Radiation," Lawrence Berkeley National Laboratory, Berckeley, 2008.

[20] Holmgren, O.; Kokkonen, K; Veijola, T.; Mattila, T.; Kaajakari, V.; Oja, A.; Knuuttila, J. V.; Kaivola, M. Analysis of vibration modes in a micromechanical square-plate resonator, *Journal of Micromechanics and Microengineering,* **2009**, 19-1, p.015028.

[21] Kaajakari, V. *Closed form expressions for RF MEMS switch actuation and release time*, *Electr. Let.*, **2009** (43)-3, pp. 149-150 DOI: 10.1049/el:20093281

[22] Volovik, G. E (1999). "Fermion zero modes on vortices in chiral superconductors". *JETP Letters* 70 (9): 609.

[23] Alastalo A. T.; Kaajakari, V.Intermodulation in capacitively coupled microelectromechanical filters, *IEEE Electron Device Letters*, **2005**, 26-, pp.189-191.

[24] Lau, J.; Lee Ch.; Premachandran C.; Aibin, Y. *Advanced MEMS Packaging* (Electronic Enginnering, Kindle Editon, USA, 2010.

Developing a Framework for Examining the Use of Mobile Transactions in Saudi Arabia: the User's Perspective

Mohammed A. Alqahtani

School of Computing Sciences
University of East Anglia, UK &
College of Computer Science and Information Technology,
University of Dammam, Saudi Arabia

Pam J. Mayhew

School of Computing Sciences,
University of East Anglia, UK

Abstract—**Both the recent advances in mobile technologies and the high penetration rate of mobile communication services have had a profound impact on our daily lives, and are beginning to offer interesting and advantageous new services. In particular, the mobile transaction (m-transaction) system has emerged, enabling users to pay for physical and digital goods and services using their mobile devices whenever they want, regardless of their location. Although it is anticipated that m-transactions will enjoy a bright future, there is apparently still reluctance among users to accept mobile transactions. Besides analyzing the literature, the authors have conducted five empirical studies to develop a robust comprehensive framework that encompasses the key factors which could affect Saudi users' intentions to use m-transactions. This paper aims to summarize and discuss these studies, show how they have evolved in several stages, aiming to reach a satisfactory level of maturity and finally shed light on interesting results.**

Keywords—Mobile Technologies; E-Commerce; M-Transactions; Conceptual Framework; Acceptance; Developing Countries

I. INTRODUCTION

Both the rapid advances in wireless technologies and the high proliferation rate of mobile communication services and artefacts have had a profound impact on industry, and are beginning to offer interesting and advantageous new services. In particular, the mobile transaction (m-transaction) system has emerged, enabling users to pay for physical and digital goods and services using their mobile devices whenever they want, regardless of their location. Nowadays, the number of mobile phones in use far exceeds any other technical devices used to market, sell, produce, or deliver products and services to consumers. They are supported, firstly, by mobile applications which have become especially valued in an era where time is precious and the weight attached to convenience is important, and secondly, by mobile communication technologies which have successfully penetrated consumer markets throughout the world. Consequently, these developments have opened up lucrative opportunities to retailers and service providers to use mobile commerce (m-commerce), and this has become a major driving force for the next phase of e-commerce [1, 2].

The importance of this study is accentuated by the facts that mobile commerce and its services are still in their infancy [3],

that there is little research addressing the acceptance of m-transactions from the user's perspective [2, 3, 4, 5, 6] and that there has been even less investigation of this issue in Saudi Arabia. The identification of factors that influence m-transaction adoption has significant value because m-transactions are likely to have a strong influence on business activities and consumer behaviour, as well as national and global markets [5, 7, 8]. There is still an apparent lack of acceptance of m-transaction services amongst consumers [6], and the success of m-transaction systems in Saudi Arabia depends on strong acceptance by mobile users and an interest to invest from both public and private stakeholders of mobile technology [1].

II. LITERATURE REVIEW

Some research firms have given strong positive predictions for the growth of mobile payment services. A management consulting firm called 'Arthur D. Little', for example, forecast a growth in mobile payment services from \$11.7 Billion in 2005 to \$37.1 Billion in 2008 [9]. This increase would have accounted for about 8% of the gross mobile services market in 2006. In spite of these hopeful forecasts, the reality seems to be quite different, and the situation is usually disappointing for those companies offering mobile payment services. According to Gartner Group [10], in 2008 only 1% of all mobile cellular users had used mobile payment services. Thus, the real market penetration of mobile payment services deviates significantly from the predictions.

Kim et al. [7] define an m-transaction as "any payment in which a mobile device is utilized to initiate, authorize, and confirm a commercial transaction". In their study they try to examine the factors influencing the intention to use m-transactions. After reviewing the literature regarding mobile payment services, and analyzing the effects of m-transaction system characteristics and consumer-centric factors on m-transaction usage across different types of m-transaction consumers, they used the Technology Acceptance Model (TAM) as a base to propose an m-transaction research model which, in addition to TAM, contains two consumer-centric factors (i.e. *personal innovativeness* and *m-payment knowledge*) and four m-transaction system characteristics (i.e. *mobility, reachability, compatibility, and convenience*). Their results show that the strongest predictors of the intention to use m-

transactions are 'perceived ease of use' and 'perceived usefulness'.

On the other hand, the literature review reveals that there is a notable shortage of research investigating the factors influencing the intention/use of m-transactions in general [3, 6], and especially in Saudi Arabia [1, 11]. This study therefore aims to contribute to closing this gap.

With regards to other research concerning m-transactions in different countries, the researchers have summarised a number of studies as follows:

TABLE I.   A NUMBER OF M-TRANSACTION RESEARCH FINDINGS

	Research Purpose	Methods Used	Theory/ Constructs Used	Results/Finding
[12]	To investigate mobile usability in mobile 'private shopping' applications.	Usability test, with a sample of 11 Turkish senior year university students.	Think aloud, eye-tracking and video recording.	Significant mobile usability problems.
[13]	Identifying the determinants of pre-adoption of mobile payment (pre-adoption and post-adoption).	Online survey. 639 respondents (483 potential adopters + 156 adopters)	Social influence; personal innovativeness; perceived risk; perceived fee; compatibility; relative advantage ➔ behavioural intention to adopt.	Behavioural beliefs in combination with social influences and personal traits are all important determinants for mobile payment services adoption and use.
[7]	To examine the factors influencing the intention to use mobile payment.	Distributed survey + email survey 269 respondents who have experience of mobile payments.	Extended TAM: perceived usefulness; perceived ease of use; personal innovativeness and m-payment knowledge; mobility, reachability, compatibility, and convenience ➔ intention to use m-payment.	The strongest predictors of the intention to use m-payment are 'perceived ease of use' and 'perceived usefulness'.
[6]	To explore the factors determining consumers' acceptance of mobile payment services.	Online survey 1447 responses (583 with experience of mobile payment services, and 864 with no experience).	Extended TAM: perceived compatibility; perceived security; perceived usefulness; perceived ease of use; individual mobility; subjective norm; attitude towards use ➔ intention to use.	Perceived compatibility, individual mobility and subjective norm have significant impact on the intention to use mobile payment services.
[14]	Developing a model which explores how customer perceptions of the value offered by m-payment services influence their attitudes towards m-payment adoption.	Distributed questionnaires 263 respondents	Service (convenience, functionality, affordability and service awareness); technology (interface, self-efficacy and security); perceived usefulness; perceived ease of use ➔ intention to use ➔ use.	Cost and individual's economic status are important decision-making factors. The familiarity and general awareness of the mobile payment service positively affect the intention to use it.
[1]	To investigate the consumers' acceptance of mobile payment in KSA.	Distributed survey 200 respondents who are mobile phone users.	Security; unauthorized use of phone to make purchases; complexity, and limited amount of money per transaction.	Security of m-payment and unauthorised use of mobile phones are the strongest concerns.
[3]	To review prior literature on mobile commerce.	Conducting an exhaustive and systematic electronic search of 59 articles (23 Chinese and 36 English).	Research subjects, research theory, and adoptive factors.	TAM is the most-used theory in consumer adoption.

III. RESEARCH METHODOLOGY

A. Choice of Approach and Methodology

Choosing an appropriate research approach during the research design is a major task for researchers [15], as there are multiple methodologies to choose from. In addition, selecting an appropriate approach is a difficult task, since the field of information systems is multi-disciplinary – "Contributions to the study of information systems come from the natural sciences, mathematics, engineering, linguistics and behavioural sciences" [16]. With regards to research methodology, the term triangulation is assigned to the practice of combining two or more research methods for the purpose of building a broader picture of the phenomenon under study. This facilitates the validation of the findings, assists in explaining diverging results and compensates for the limitations inherent in a single research method. Triangulation is defined by Cohen, Manion and Morrison [17] as "the use of two or more methods of data collection in the study of some aspects of human behaviour".

On the other hand, Creswell and Clark [18] highlight the importance of mixed methods when one data source may be insufficient, and for the following additional reasons: (1) to explain initial results; (2) to generalize exploratory findings; (3) to enhance a study with a second method; (4) to best employ a theoretical stance and (5) to understand a research objective through multiple research phases. Furthermore, diversity in research methodologies (i.e. qualitative and quantitative) is considered as a major strength of information systems (IS) research. Mixing methodologies in research is considered potentially superior to a single method design [19].

Bryman and Bell [20] explain in their book the difference between quantitative and qualitative methods. Quantitative research can be interpreted as a research strategy that stresses quantification in collecting and analysing the data, involving a deductive approach to the relationship between theory and research, in which the accent is placed on the testing of theories. It has incorporated the practices and norms of the natural scientific model and of positivism in particular; and represents a view of social reality as an external objective reality. On the other hand, qualitative research can be interpreted as a research strategy that usually places greater emphasis on words rather than quantification in collecting and analysing the data, involving an inductive approach to the relationship between theory and research, in which the emphasis is placed on the formation of theories. It has rejected the practices and norms of the natural scientific model and of positivism in particular, in preference for an emphasis on the ways in which individuals construe their social world and represents a view of social reality as a constantly shifting emergent property of individuals' creation [20].

In this research, it was decided to employ a blend of quantitative and qualitative methods as well as triangulation of techniques in order to get a wider picture of the phenomenon under study (i.e. the intention to use m-transactions from Saudi users' perspective). It was decided to employ triangulation as mentioned earlier, integrating the following methods: *Descriptive/interpretive, Interviews, Surveys, Focus Group,* and *'Think-aloud' as an evaluation method (user usability test)*

to study this area. The following table (Table II) summaries the research approach, methodologies and data collection.

TABLE II.   SUMMARY OF RESEARCH APPROACH AND METHODOLOGY

No	Methodology	Users	Literature	Tool	Sample Size
1	Qualitative analysis	☒	☑	Descriptive/ Interpretive	-
2	Qualitative analysis	☑	☑	Exploratory (1) Interviews	40
3	Qualitative analysis	☑	☑	Exploratory (2) Interviews	122
4	Qualitative analysis	☑	☑	Exploratory (3) Focus Group	8
5	Quantitative analysis	☑	☑	Validation (1): Questionnaire	1008
6	Qualitative analysis	☑	☑	Validation (2): Think-Aloud (Usability test)	30

B. Descriptive/Interpretive

Descriptive or interpretive research usually refers to the existing literature or past achievements, in addition to actual present happenings. Hart [21] defined a literature review as "the use of ideas in the literature to justify the particular approach to the topic, the selection of methods, and demonstration that this research contributes something new". A methodological review of past literature is a critical endeavour for any academic research [22] and the need to discover what is already known in the body of knowledge prior to introducing any research study should not be underestimated [21]. Significant developments in our knowledge, and in our ability to develop theories, can be achieved through an in-depth review of this kind. A comprehensive review of past research/developments may not only lead to new visions but is also more likely to ensure that subsequent research is based on past endeavours. Researchers, furthermore, have criticized the Information Systems (IS) field for having insufficient theories and outlets for a quality literature review [23]. Moreover, Webster and Watson [22] noted that the IS field should greatly benefit from effective methodological literature reviews that are "… strengthening IS as a field of study". The strengths of this method of research lie in its capability to represent reality following an in-depth self-validating process in which assumptions/presuppositions are constantly questioned, and the understanding of the phenomena under study is refined. The weaknesses of this approach may include the problems reviewers face in understanding the results of a research with which they may be unfamiliar. Other influential matters are associated with the researchers' skills and ability to recognize their biases and assumptions. In this research, this approach was employed to review the literature related to the subject in hand.

C. Exploratory Study (1) Interviews

The reason for using a qualitative method in this research is that it enables the researchers to reach deeper into the experience of the participants, to find out how opinions and habits are shaped culturally in order to discover the relevant variables. In their book, Corbin and Strauss [24] summarize the benefits of qualitative research as sharing these characteristics: 'a humanistic bent', 'curiosity', 'creativity and imagination', 'a sense of logic', 'the ability to recognize diversity as well as

regularity', 'a willingness to take risks', 'the ability to live with ambiguity', 'the ability to work through problems in the field', 'an acceptance of the self as a research instrument', and 'trust in the self and the ability to see value in the work that is produced'. This research has been conducted using a grounded theory methodology (GTM) which was developed by Glaser and Strauss in 1967. Following GTM allows a qualitative method to be used which provides the following valuable characteristics: shedding light on a person's daily life experience; evaluating contributors' perspectives; investigating the interactive processes between researcher and respondents; and being descriptive based on people's words [25].

This research [26] aimed to elicit consumers' opinions i.e. what people in Saudi Arabia think and believe about adopting e-commerce. This included thoughts, beliefs and opinions. This information was gathered by asking general questions and recording notes and comments about the important ideas and concepts. The semi-structured interviews were conducted during August 2011 in Dammam and Khobar, as they are two of the main cities in Saudi Arabia, which are highly populated and have high business potential. All the participants were selected at random regardless of their age, income and gender. The researchers approached people from private and public organizations in different locations (e.g. universities, companies, schools, Internet cafes, streets and shops). However, this research is about adopting e-commerce, so the researchers had to verify that the participants were Internet users before starting the interviews. The sample size was 40 interviewees.

D. Exploratory Study (2) Interviews

A semi-structured interview instrument was developed in which all questions were based on the study model, a model which was an adopted version of Unified theory of acceptance and use of technology UTAUT. As before, this research [11] aimed to discover consumers' opinions about adopting and using m-transactions, and this information was gathered by asking general questions and recording notes and comments about the important ideas and concepts. For validity reasons, the instrument was extensively pre-tested and evaluated by academic and practical experts to whom copies of the interview's questions were sent for judgment. The instrument was then piloted and adjustments were made accordingly with the support of professional experts, with particular attention being given to the wording and the overall structure and presentation of the interview items. The semi-structured interviews were conducted during May/June 2012 in Saudi Arabia. The researchers aimed to interview the public with no limit on special features or demographic characteristics; hence all participants were selected at random regardless of their age, income and gender. The researchers approached people from private and public organizations in different locations as before (e.g. universities, companies, schools, Internet cafes, streets and shops), and again verified that the participants were Internet users before starting the interviews. Importantly, the interviewer clearly informed the interviewees that any data gathered would be kept anonymously and that the interviewees had the option to receive an executive summary of the study results. Furthermore, the authors aimed to record the interviews in electronic form to ease the later analysis. The sample size of these interviews was 122 interviewees. The reliability of the

instrument was scrutinized by making sure that there were not many gaps in the collected data between different respondents. This study initially involved exploratory research using a qualitative approach. By adopting this approach the researchers were able to gain an in-depth understanding of consumers' concerns regarding adopting m-transactions in Saudi Arabia.

E. Exploratory Study (3) Focus Group

In terms of explorative studies, focus group interviews have been suggested as a suitable method [27] and previous research has highlighted their feasibility and capability for studying innovative mobile transactions [28, 29]. The strength of the focus group interview method is that it is dynamic and interactive, so it has the ability to provide researchers with elaborated perspectives of the participants on the topic under discussion; it has been considered an especially informative way of developing a research model in a new research area [29, 30]. For the focus group members to interact successfully and work dynamically as a group, a number of important criteria need to be met in the selection of members. Stewart and Shamdasani (1990, p. 33) note that "the usefulness and validity of focus group data are affected by the extent to which participants feel comfortable about openly communicating their ideas, views or opinions" [cited in 29]. Groups that have formed naturally have proved to be particularly relaxed, thus easing the conversations amongst the participants [31]. As a result, in order to guarantee a proper discussion and interaction throughout the sessions, a naturally formed group was selected for this study with a total of 8 mobile experts. The participants knew each other as classmates, friends, co-workers or via a common hobby. Experience in online purchasing and using a mobile phone that has internet access were two factors estimated to be necessary in order for the participants to be able to discuss the relevant topic. According to Krueger and Casey [32] the recommended size for a focus group ranges from 4 to 12, therefore 8 participants were involved in this study. The majority held a postgraduate degree, and most of the participants (6 out 8) had experience of mobile transactions.

F. Validation Study (1) Questionnaire

The variables in the developed framework were latent and could not be measured directly. Thus, a set of measurement indicators was generated to operationalize each construct, using existing indicators from previous studies when available, or adapting them if necessary. All constructs were measured reflectively. Each of the indicators was measured using a 7-point Likert-scale. These interval scales ranged from 1 – strongly disagree, 2 – disagree, 3 – disagree somewhat, 4 – not applicable, 5 – agree somewhat, 6 – agree, and 7 – strongly agree. The survey data, with regards to demographics and descriptive data, was primarily measured using nominal scales. Using the empirical data from the distributed questionnaires, the measurement properties were assessed and hypotheses were verified using the partial least squares structural equation modelling (PLS-SEM) approach [33, 34]. The PLS-SEM approach has enjoyed steady popularity as a key multivariate analysis method in management information system (MIS) research [35, 36]. Structural equation models (SEM) allow both exploratory and confirmatory modelling, meaning that they are suited to both theory development and theory testing. Confirmatory modelling usually starts with a hypothesis that is represented in a causal model. The concepts used in the model should then be operationalized to allow testing of the relationships between the concepts. The model is tested against the obtained measurement data to identify how well the model fits the data. The causal assumptions embedded in the model frequently have falsifiable implications which can be tested against the data. [37].

This approach (PLS-SEM) was chosen for the data analysis since, compared to covariance-based approaches, it is advantageous when the research model has large numbers of indicators and is relatively complex, when the measures are not well established, and/or the relationships between the indicators and latent variables may need to be modelled in different modes (e.g. formative and reflective measures) [38, 39, 40]. Furthermore, PLS may be better suited as it has fewer demands with regards to sample size and residual distributions [35, 39, 40, 41]. The software package SmartPLS [42] was used for the statistical calculations. A PLS path model consists of two elements. First, there is a structural model (also called the inner model in the context of PLS-SEM) that represents the relationships (paths) between the constructs. Second, there are the measurement models (also referred to as the outer models in PLS-SEM) of the constructs that display the relationships between the constructs and the indicator variables. In general there are two types of measurement models: one for the exogenous latent variables (i.e. those constructs that explain other constructs in the model) and one for the endogenous latent variables (i.e. those constructs that are being explained in the model) [43].

As mentioned above, model estimation delivers empirical measures of the relationships between the indicators and the constructs (measurement models), as well as between the constructs (structural model). The empirical measures enable us to compare the theoretically established measurement and structural models with reality, as represented by the sample data. In other words, we can determine how well the theory fits the data.

G. Validation Study (2) Think-Aloud (Usability test)

Choosing the correct evaluation method is important; scientifically validated information on appropriate testing methods is valuable for usability practitioners. The Thinking-Aloud protocol is one of the most important usability evaluation methods (UEMs). It has been a key evaluation framework ever since the 1980s, and it remains important today in the Information System (IS) field [44, 45, 46]. In this study, user testing and think-aloud 'concurrent approach' methods were chosen. These methods are the best way to examine whether usability is indeed the critical factor that affects Saudi Arabian consumers' intention to use m-commerce.

The experiments took place in a typical usability lab setting in Saudi Arabia from July 2014 to September 2014. As in previous studies, all the participants were selected at random regardless of their age, income and gender. The researchers approached people from private and public organizations in different locations (e.g. universities, companies, schools, Internet cafes, streets and shops). As this research is about the usability of m-transactions on the Souq.com app, the researchers had to verify that the participants were mobile

users, familiar with mobile commerce, and had conducted a mobile transaction before, but had not used the tested application before (this was verified in the demographic section of the questionnaire). The sample size of this experiment was 30. One of the researchers (the main author) adopted the role of the evaluator in all the sessions, noting all the comments made by the users. The evaluator researcher used a stopwatch to record the time spent by each user on each task, and an observation sheet to write down the behaviour of each user and the problems discovered. After the users had finished their tasks, they were asked to rate their level of satisfaction in a System usability scale (SUS) questionnaire with a five point scale – 1 for strongly disagree and 5 for strongly agree. Bangor, Kortum and Miller [47] suggested slightly modified statements from those used in the original SUS instrument by [48]. The changes are (1) replacing the word "cumbersome" with "awkward" and (2) replacing the word "system" with "product". These studies applied this slightly modified SUS and also replaced the word "product" with "application". To add a valid assessment of what the absolute numerical score of SUS means, a 7 point adjective rating scale can be used to judge the SUS scores. This has become known as the "university grade analog" [47]. This study applied this technique and again replaced the word "product" with "application". Each participant conducted the experiment individually and the researcher ensured that every participant felt at ease whilst performing the tasks. Furthermore, users were given a chance to fill in an open-ended questionnaire by writing down their comments and feedback on the application, and explaining any reaction that was observed during the test. Subsequently, the researcher extracted the problems from the observer sheet and they were merged into a final master problem list.

IV. RESULTS AND DISCUSSION

A. Exploratory Study (1) Interviews

The findings of this study can be divided into two main categories that both play a significant role, from a consumer perspective, in the adoption of e-commerce. The first is 'Enablers', which is comprised of 10 factors: Internet and Infrastructure, security, trust, cyber law, awareness and perceived usefulness, postal services, government e-readiness, the presence of commercial electronic websites, and warranty. Some of these need more improvement and expansion, supported by current technologies, and some of them need to be encouraged and widely applied using the best and most advanced technologies. What is required are reliable internet connections; good infrastructure; the understanding and implementation of cyber law; efficient transportation for the delivery of products; trustworthy security systems; fair and effective warranty agreements; the perception of trust, and the promotion of it; comprehensive awareness and training programs; and e-readiness and support from government. This study also identified 'Disablers', which is comprised of 5 factors as follows: fraud and hacking, resistance to change, Telecom charges (cost), tangibility, trial and experience. These factors need to be either managed, limited or prevented in the following ways: finding alternative ways of offering customers a trial or the equivalent of a tangible experience of the product; introducing competitive and affordable charges; promoting the

potential benefits of e-commerce to eliminate the impact of resistance to change; and installing high quality security systems using strong encryption algorithms to prevent hacking and fraud.

B. Exploratory Study (2) Interviews

In this study, based on the literature and the collected consumers' feedback, the researchers divided the different items (factors) into four main categories that play a significant role in the acceptance or use of m-transactions. These are described below. (1) 'Design and language support', which includes 2 factors: 'usability' and 'Arabic language support', both of which need to be encouraged, improved and widely applied using the best and most advanced technologies. What is required here is that mobile websites are very user-friendly when navigating them; that there are reliable internet and wireless connections, and a very wide coverage for the reception of new mobile technologies (e.g. 3G and 4G); websites should be fully supported by a reasonable range of different mobile devices, and they should fully support the Arabic language, especially for conducting financial transactions. (2) 'User-related Factors', which encompasses two factors: 'trust' and 'awareness'. Trust needs to be fostered and signs need to be used that can encourage more trust (e.g. a trusted third party sign); comprehensive awareness and training programs are required to increase the level of awareness amongst Saudi people. (3) 'Organizational-related Factors' which includes six factors: 'telecommunication infrastructure', 'postal services', 'government e-readiness', 'cyber law', 'telecom charges' (cost), and 'payment gateways'. These require a sophisticated and reliable wireless telecommunication infrastructure, the understanding and implementation of cyber laws, efficient transportation for the delivery of products and e-readiness, and support from government, introducing competitive and affordable charges, and providing a variety of different payment methods. (4) 'System-related Factors', which contains 'security, hacking and fraud', 'usefulness' and 'privacy'. This category requires the installation of high quality, trustworthy security systems that use strong encryption algorithms to prevent hacking and fraud, and therefore increase privacy.

C. Exploratory Study (3) Focus Group

All the results from the previous studies by the authors [11, 26] were examined in this focus group study. The results that emerged from the discussions in the group were organized into 11 key factors which are: ease of use, visual appeal, navigational structure, ICT infrastructure, usefulness, cost, government m-readiness, social influence, security, trust and culture. Other aspects were mentioned in the discussion, such as: awareness, payment gateway, m-commerce diffusion, etc., but these concepts were considered as insignificant or irrelevant by the participants during the discussion.

D. Validation Study (1) Questionnaire

The developed framework was empirically validated in a study which involved a statistically representative sample size of more than 1,000 Saudi users from different demographic backgrounds. The empirical analysis revealed that security, ease of use, individualism, masculinity, navigational structure, power distance (strength of social/employment hierarchy),

uncertainty avoidance, usefulness, and website support for mobiles all have a significant impact on consumers' intentions to use m-transactions. Amongst these factors, ease of use was the most influential. This study has been started in 2014 and the authors aim to publish its results as soon as possible. The research results showed that the usability of m-transactions was the most important concern for Saudi users [11, 26, 49]. In particular this study aimed to validate this result by conducting a usability test for m-transactions with Saudi users. The feedback from these tests should help to elucidate the results and shed some light on why usability was ranked higher than other factors.

E. Validation Study (2) Think-Aloud (Usability test)

Although the tested application Souq.com [50] enjoys great popularity and represents realistic m-transaction functionalities, the results clearly showed that it was unsatisfactory as a usable application. It has a very low SUS Score (50) compared to 60 which is recommended as an acceptable score. It also ranked in Bangor, Kortum and Miller [51] classification as 'F' which is considered as unacceptable. Furthermore, it has a very low success rate and efficiency. Therefore, the designers and developers of this application have to pay more attention to the usability level of their application and consider the usability problems discovered through the experiments, especially to those categories with a higher number of problems such as 'design/layout', 'completeness', 'correctness' and 'comprehension'. In more detail, the problems and issues that need to be considered are: (1) simplifying the forms (e.g. registration, contact us and add an address); (2) clearly labelling the buttons with names that are easy to understand and reflect the actual function of that button; (3) redesigning the filter function in all sections in a simple way (e.g. so there is no need to press 'apply' for every specification for finding a product); (4) managing the basket needs to be improved (e.g. the product amount does not handle the value of 0 (zero) and the button 'add to the basket' needs to be renamed); (5) adding further functions (e.g. the ability to delete more than one item at a time); (6) correcting errors of logic (e.g. the authors' names were written in two different languages, English and Arabic, the challenge question was written in Arabic yet does not accept the answer while the keyboard is Arabic, and the whole payment confirmation page was written in English while the chosen interface language was Arabic); and (7) removing the distracting popup messages (e.g. after clicking contact us or while looking for a product). A usability test should be conducted prior to the launching of any mobile application to ensure that the usability level is satisfactory. Future research has to pay attention to the importance of usability to consumers' acceptance of IT innovations such as m-transactions.

V. Conclusion

The ultimate product of this research is to develop a consolidated framework for the intention to use m-transactions, combined with a set of recommendations for mobile websites and application developers, designers, government, and organizations which intend to extend their business into the mobile commerce area, and eventually the users. In order to achieve this goal, this study evolved in several stages aiming to reach a satisfactory level of maturity. These stages can be divided into three main phases: three exploratory studies, the first of which (1) shed light on e-commerce as the first generation of m-commerce, while the other two studies focused on m-transactions. These studies helped to add the cultural qualities as a further dimension that would play a significant role in such a unique cultural region. Consequently, a holistic framework is integrated that includes the key factors affecting the intention to use m-transactions. This framework is empirically validated in (2) a further study using a statistically representative sample size of about 1000 Saudi users from different demographic backgrounds. The empirical analysis revealed that security, ease of use, individualism, masculinity, navigational structure, power distance, uncertainty avoidance, usefulness, and website support for mobiles have a significant impact on the intention to use m-transactions. Amongst those factors, ease of use was the most influential for the intention to use m-transaction. Therefore, this lead to (3) another study aimed to empirically investigate the level of ease of use (usability) of conducting m-transaction within the Saudi context. In total, this research went through five different empirical studies to extend our understanding of the phenomenon of m-transaction. Please see Figure 1 below, which summarizes all the different stages of the research.

Fig. 1. A summary of the studies' results and the developed conceptual framework

REFERENCES

[1] O. Bamasak, 'Exploring consumers acceptance of mobile payments – an empirical study'. *Int J Information Technology, Communications and Convergence,* 1, 173-85, 2011.

[2] T. Dahlberg, Mallat N., Ondrus J. & Zmijewska A., 'Past, present and future of mobile payments research: A literature review'. *Electronic Commerce Research and Applications,* 7, 165-81, 2008.

[3] G. Wei, Xinyan Z. & Yue M., 'Literature review on consumer adoption behavior of mobile commerce services'. *E -Business and E -Government (ICEE), 2011 International Conference on*, 1-5, 2011.

[4] H. Huang, Liu L. & Wang J., 'Diffusion of Mobile Commerce Application in the Market'. *Innovative Computing, Information and Control, 2007 ICICIC '07 Second International Conference on*, 485-, 2007.

[5] H. Dai & Palvia P. C., 'Mobile Commerce Adoption in China and the United States: A Cross-Cultural Study'. *SIGMIS Database,* 40, 43-61, 2009.

[6] P. G. Schierz, Schilke O. & Wirtz B. W., 'Understanding consumer acceptance of mobile payment services: An empirical analysis'. *Electronic Commerce Research and Applications,* 9, 209-16, 2010.

[7] C. Kim, Mirusmonov M. & Lee I., 'An empirical examination of factors influencing the intention to use mobile payment'. *Computers in Human Behavior,* 26, 310-22, 2010.

[8] Y. A. Au & Kauffman R. J., 'The economics of mobile payments: Understanding stakeholder issues for an emerging financial technology application'. *Electron Commer Rec Appl,* 7, 141-64, 2008.

[9] Arthur D. Little. Global M-Payment Report 2004. Vienna: 2004.

[10] Gartner Inc. Gartner Group Dataquest Insight: Mobile Payment, 2007–2012. Stamford, CT: 2009.

[11] M. Alqahtani, Al-Badi A. & Mayhew P., 'Exploratory Study of M-Transaction: User's Perspectives'. *The Electronic Journal of Information Systems in Developing Countries,* 60, 1-22, 2014.

[12] Ö. Öztürk & Rızvanoğlu K., M-Commerce usability: an explorative study on turkish private shopping apps and mobile sites. *Design, User Experience, and Usability Web, Mobile, and Product Design.* City: Springer, 2013.

[13] S. Yang, Lu Y., Gupta S., Cao Y. & Zhang R., 'Mobile payment services adoption across time: An empirical study of the effects of behavioral beliefs, social influences, and personal traits'. *Computers in Human Behavior,* 28, 129-42, 2012.

[14] K. Petrova & Mehra R., "Mobile Payment: An Exploratory Study of Customer Attitudes". In*: Wireless and Mobile Communications (ICWMC), 2010 6th International Conference on, 20-25 Sept. 2010* 2010. 378-83, Year.

[15] G. Walsham, 'The emergence of interpretivism in IS research'. *Information systems research,* 6, 376-94, 1995.

[16] F. Land, 'The information systems domain'. *Information systems research: issues, methods and practical guidelines*, 6-13, 1992.

[17] L. Cohen, Manion L. & Morrison K., 'Research Methods in Education [5 th edn] London: Routledge Falmer'. *Teaching in Higher Education,* 41, 2000.

[18] J. W. Creswell & Clark V. L. P. *Designing and Conducting Mixed Methods Research*, SAGE Publications, 2011.

[19] V. Venkatesh, Brown S. A. & Bala H., 'Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems'. *Mis Quarterly,* 37, 21-54, 2013.

[20] A. Bryman & Bell E. *Business Research Methods*, Oxford University Press, 2007.

[21] C. Hart. *Doing a literature review: Releasing the social science research imagination*, Sage, 1998.

[22] J. Webster & Watson R. T., 'Analyzing the past to prepare for the future: Writing a literature review'. *Management Information Systems Quarterly,* 26, 3, 2002.

[23] Y. Levy & Ellis T. J., 'A systems approach to conduct an effective literature review in support of information systems research'. *Informing Science: International Journal of an Emerging Transdiscipline,* 9, 181-212, 2006.

[24] J. Corbin & Strauss A. *Basics of qualitative research:Techniques and procedures for developing grounded theory, 3rd ed., 3*, Thousand Oaks, Calif. ; London : SAGE Publications., 2008.

[25] C. Marshall & Rossman G. B. *Designing Qualitative Research, 3rd ed.*, Thousand Oaks, Ca., Sage., 1999.

[26] M. Alqahtani, Al-Badi A. & Mayhew P., 'The Enablers and Disablers of E-Commerce: Consumers' Perspectives'. *The Electronic Journal of Information Systems in Developing Countries EJISDC,* 54, 1-24, 2012.

[27] B. J. Calder, 'Focus Groups and the Nature of Qualitative Marketing Research'. *Journal of Marketing Research,* 14, 353-64, 1977.

[28] S. L. Jarvenpaa & Lang K. R., 'Managing the Paradoxes of Mobile Technology'. *Information Systems Management,* 22, 7-23, 2005.

[29] N. Mallat, 'Exploring consumer adoption of mobile payments – A qualitative study'. *The Journal of Strategic Information Systems,* 16, 413-32, 2007.

[30] S. Wilkinson, Focus group research. *In:* Silverman D. (ed.) *Qualitative Research: Theory, Method and Practice.* Second ed. City: SAGE, 2004.

[31] A. Bryman. *Social Research Methods, second ed.,* New York, Oxford University Press, 2004.

[32] P. R. A. Krueger & Casey M. A. *Focus Groups: A Practical Guide for Applied Research,* Thousand Oaks, Sage Publications, 2009.

[33] W. W. Chin, Commentary: Issues and opinion on structural equation modeling. JSTOR, 1998.

[34] H. Wold, 'Soft modelling: the basic design and some extensions'. *Systems under indirect observation, Part II*, 36-7, 1982.

[35] C. M. Ringle, Sarstedt M. & Straub D. W., 'Editor's comments: a critical look at the use of PLS-SEM in MIS quarterly'. *MIS quarterly,* 36, iii-xiv, 2012.

[36] D. Gefen, Rigdon E. E. & Straub D. W., 'Editor's Comment: An Update and Extension to SEM Guidelines for Administrative and Social Science Research'. *MIS Quarterly,* 35, iii-xiv, 2011.

[37] K. A. Bollen. *Structural equation models*, Wiley Online Library, 1998.

[38] W. W. Chin & Newsted P. R., 'Structural equation modeling analysis with small samples using partial least squares'. *Statistical strategies for small sample research,* 1, 307-41, 1999.

[39] C. Fornell & Bookstein F. L., 'Two structural equation models: LISREL and PLS applied to consumer exit-voice theory'. *Journal of Marketing research*, 440-52, 1982.

[40] N. Urbach, Smolnik S. & Riempp G., 'An empirical investigation of employee portal success'. *The Journal of Strategic Information Systems,* 19, 184-206, 2010.

[41] D. Gefen, Straub D. & Boudreau M.-C., 'Structural equation modeling and regression: Guidelines for research practice'. *Communications of the association for information systems,* 4, 7, 2000.

[42] C. M. Ringle, Wende S. & Becker J.-M. 2014. *Smartpls 2.0* [Online]. City: University of Hamburg, Hamburg. Available: http://www.smartpls.com [Accessed 02-10 2014].

[43] J. F. Hair Jr, Hult G. T. M., Ringle C. & Sarstedt M. *A primer on partial least squares structural equation modeling (PLS-SEM)*, SAGE Publications, Incorporated, 2014.

[44] O. Alhadreti, Alroobaea R., Wnuk K. & Mayhew P. J., "The Impact of Usability of Online Library Catalogues on the User Performance". In, 2014 2014. IEEE, 1-4, Year.

[45] M. J. Van Den Haak, De Jong M. D. T. & Schellens P. J., 'Evaluating municipal websites: A methodological comparison of three think-aloud variants'. *Government Information Quarterly,* 26, 193-202, 2009.

[46] R. Benbunan-Fich, 'Using protocol analysis to evaluate the usability of a commercial web site'. *Information & Management,* 39, 151-63, 2001.

[47] A. Bangor, Kortum P. T. & Miller J. T., 'An empirical evaluation of the system usability scale'. *Intl Journal of Human–Computer Interaction,* 24, 574-94, 2008.

[48] J. Brooke, 'SUS-A quick and dirty usability scale'. *Usability evaluation in industry,* 189, 194, 1996.

[49] M. Alqahtani, Alroobaea R. S. & Mayhew P. J., "Building a Conceptual Framework for Mobile Transaction in Saudi Arabia: A User's Perspective". In*:* Science and Information Conference (SAI), 2014 London. 967-73, 2014.

[50] Souq.Com. 2014. *About us* [Online]. City: Saudi.souq.com. Available: http://saudi.souq.com/sa-en/about-us/c/ [Accessed 03/09/2014 2014].

[51] A. Bangor, Kortum P. & Miller J., 'Determining what individual SUS scores mean: Adding an adjective rating scale'. *Journal of usability studies,* 4, 114-23, 2009.

A Safety Analysis Approach to Clinical Workflows: Application and Evaluation

Safety Analysis of Clinical Workflows

Lamis Al-Qora'n, Neil Gordon, Martin Walker, Septavera Sharvia, Sohag Kabir
Department of Computer Science
University of Hull
Hull, United Kingdom

Abstract—**Clinical workflows are safety critical workflows as they have the potential to cause harm or death to patients. Their safety needs to be considered as early as possible in the development process. Effective safety analysis methods are required to ensure the safety of these high-risk workflows, because errors that may happen through routine workflow could propagate within the workflow to result in harmful failures of the system's output. This paper shows how to apply an approach for safety analysis of clinical workflows to analyse the safety of the workflow within a radiology department and evaluates the approach in terms of usability and benefits. The outcomes of using this approach include identification of the root causes of hazardous workflow failures that may put patients' lives at risk. We show that the approach is applicable to this area of healthcare and is able to present added value through the detailed information on possible failures, of both their causes and effects; therefore, it has the potential to improve the safety of radiology and other clinical workflows.**

Keywords—*clinical workflows; safety analysis; radiology; HiP-HOPS*

I. INTRODUCTION

Clinical workflow as defined by [1] is a term that is used to describe the healthcare activities that are performed carefully by more than one member to accomplish a clinical process (e.g. treatment or diagnosis) and to produce a certain clinical service.

Due to the growing number of adverse events, risk management of healthcare activities, the issue of patient safety, medical errors prevention and adverse events reporting are broadly studied nowadays. A report in 1999 entitled "To Err is Human: Building a Safer Health System" which was released by the Institute of Medicine (IOM) stated that errors cause between 44000 and 98000 deaths every year in American hospitals and over one million injuries [1]. Moreover, around 425,000 patients (5% of total) admitted to hospitals in England and Wales each year experience adverse events from medical errors [2].

So, clinical workflows can be described as safety critical workflows because they have the potential to cause harm or death to patients. Their safety needs to be considered as early as possible in their development process, where the safety analysis results can be used to refine the models and to derive more detailed functional models and specifications of the workflow.

In [3] an approach to safety analysis of clinical workflows was proposed, which is explained in the next section. The approach is applied to a case study and an evaluation of the approach and its benefits is shown through the paper.

II. AN APPROACH FOR SAFETY ANALAYSIS OF CLINICAL WORKFLOWS

The following figure shows an approach for safety analysis of clinical workflows proposed by [3]:
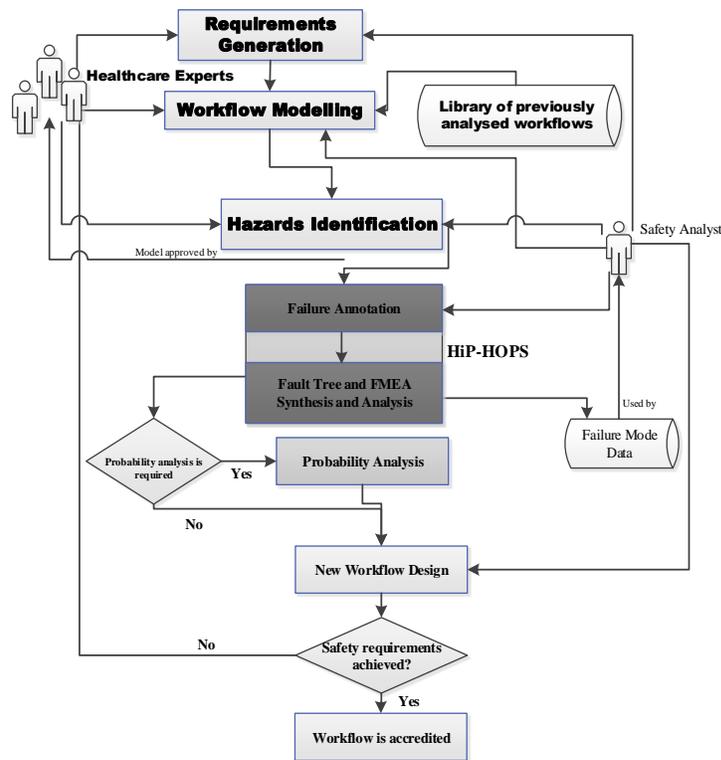
Fig.1. An Approach to Safety Analysis of Clinical Workflows [3]

The approach is designed to support the development and safety analysis of clinical workflows. It starts with the process of requirements generation where the workflow is documented in order to understand it, then the workflow documentation is translated into models by the safety analyst and the healthcare team. After that the safety analyst - in cooperation with the healthcare team - start the process of hazards identification in

order to extend the models with the local behavior of each component, and then in an automatic manner the tool (Hierarchically-Performed Hazard Origin and Propagation Studies (HiP-HOPS)) generates both Fault Tree Analysis (FTA) and Failure Mode and Effect Analysis (FMEA) results. These results are qualitative in the sense that they show how the failure of a single component or combinations of failures of different components can lead to system failure. If the results of qualitative analyses are enough then the new workflow design can be generated based on the outcomes of the analyses. Otherwise, if quantitative analysis is required then it has to be done before starting with the new workflow design. Finally, if the workflow requirements are achieved then the workflow can be accredited.

FTA [4] is a common safety analysis technique through which root causes of an undesired event are identified. It is a deductive technique which determines how an undesired event (often termed the *top event*) can be caused by lower level failures (or events) or their combinations. Quantitative analysis of the FTA can be implemented to calculate the probability of the top event and qualitative analysis is performed to identify the necessary and sufficient combinations of events which can cause the top event (termed Minimal Cut Sets (MCS)). Quantitative analysis of a fault tree, which follows qualitative analysis, can help to estimate the probability of the top event occurring from the given failure rates of basic failure modes of the system. Failure Mode and Effect Analysis (FMEA), on the other hand, is an inductive safety analysis technique that examines the effect of lower level (component) failures towards the higher-level system failures. FTA and FMEA has a wide use in exploring and analysing healthcare issues related to patient safety (e.g. [5]; [6]; [7]), and they showed their ability to analyse clinical processes. Automated FTA and FMEA would present and provide more efficiency in analysing clinical processes.

HiP-HOPS which was initially proposed by [8], is a state-of-the-art technique, which has been prominently used in mechanical systems to effectively identify weak points in system design. It is a predictive safety analysis technique which enables semi-automated FTA and FMEA. In other words, it incorporates, automates, and integrates a number of classical techniques. The current implementation of the HiP-HOPS has the design optimisation capability that can help to select component and subsystem among different alternatives as well as helps to decide the level and location of replicated components.

HiP-HOPS works in combination with a number of frequently used system modelling tools or packages (e.g. Matlab Simulink), from which it receives block diagrams of systems being analysed and associated failure behaviour. It includes three main phases: a modelling phase, a synthesis phase, and an analysis phase where MCSs and FMEA are generated. The process starts when designers build a model of the system, then they annotate the model and its components with detailed failure information. Internal failure information can be annotated into the components as a set of Boolean expressions that are manually added to each component to describe how failures of the component output can be caused

by a single input failure and/or a combination of input failures and/or by internal malfunctions of the component itself. These expressions essentially represent the component fault trees of the system components describing the generation, propagation, and transformation of failures between the inputs and outputs of the components. After defining the behaviour of a component, the component can be stored in the library to allow greater degree of reusability.

Qualitative analysis is performed based on the logical failure behaviour of the components and it starts with a top event (system failure) and traverses the model by following the propagation of the failures backwards from the top level of the system towards the basic component level. The outcome of this process is a fault tree showing the failure behaviour of the whole system. As this fault tree is relatively complex therefore it is minimised by applying logical rules to obtain minimal cut sets. As part of the qualitative analysis, FMEA is also generated from the fault trees.

In addition to the logical failure behaviour of the components, numerical data (e.g. failure rate, severity of the component) can also be entered for the components. Quantitative analysis can be performed based on the numerical data entered for the components. As seen in Figure 1, the quantitative analysis is optional in the proposed approach. However, if quantitative analysis is required and sufficient data for the analysis are available then it is possible to quantify the fault tree to get the probability of the top event. As MCSs of the fault tree are represented as the conjunction (AND gate) of the statistically independent basic failure modes therefore the probability of a MCS be obtained using the following equation.

$$P(MCS_i) = \prod_{j=1}^{n} P(BE_j) \qquad (1)$$

Where $P(MCS_i)$ is the probability of the minimal cut set *i* and $P(BE_j)$ is the probability of the basic event *j*.

Since the top event is represented as the disjunction (OR gate) of the MCSs, therefore, the top event probability can be calculated as [9]:

$$P(top\ event) = 1 - \prod_{i=1}^{n}(1 - P(MCS_i)) \qquad (2)$$

where $P(top\ event)$ is the probability of the top event.

In the clinical workflows, a lot of human activities are involved; therefore, human errors are expected to constitute a great proportion of the basic failure modes. It is relatively easy to quantify the failure probability of mechanical components; in contrast, it is difficult to quantify the probability of the human error due to the uncertainty involved in quantifying human behaviour. So, uncertainty in human behaviour may require to be considered in the quantitative analysis. One possible way is to translate the fault tree into Bayesian Networks (BNs) using the method shown in [10] and then perform the probabilistic analysis because BNs are considered as efficient methods for performing probabilistic inference under uncertainty.

HiP-HOPS can in general be applied to systems that involve data, information or material flow. However, in our case "components" may represent clinical processes, humans, tasks, or any other components of a clinical workflow architecture.

HiP-HOPS was proposed by [11] to analyse the safety of the workflow of a home Telemonitoring system. This paper applies an integrated approach which utilises HiP-HOPS to conduct safety analysis of a RIS/PACS workflow. The result of the analysis is the root causes of different failures, and their direct and indirect effects on both the workflow and the patients themselves.

III. APPLICATION OF THE APPROACH TO A CASE STUDY

Radiology Information Systems (RIS) and Picture Archiving and Communication Systems (PACS) technology has advanced dramatically in recent years, including the technology of acquiring, storing, retrieving, displaying, and distributing clinical images [12]. It has become a mature technology and has been commonly implemented in a number of developed countries [13]. Different systems have been designed and developed to assist different workflows in the radiology departments in several hospitals. In Jordan for example, RIS/PACS are implemented in a number of private, government, and military hospitals. To investigate the concerns that medical staff have due to the adoption of RIS/PACS systems, we conducted a number of interviews in one of the Jordanian hospitals. These were followed by another set of interviews to document the workflow in the radiology department in the same hospital, where RIS/PACS has been adopted. We found that faults and errors in the workflows might lead to harmful failures in the outputs (e.g. producing a report that has an incorrect description of the patient's situation, or leading to undesired reactions by the patient). Having the wrong report potentially results in an incorrect diagnosis and treatment, placing the patient's life at risk, while the effect of having unwanted side effects by the patient varies depending on how serious these effects are.

With this prevalence of RIS/PACS in healthcare institutions, there is a growing need to analyse their workflow safety, both ensuring the safety of the workflow design and the safety during the operational phase. In other words, securing the design of the theoretical workflow in terms of safety issues, and then making sure about following this workflow in the operational phase. Analysing and modelling the workflow plays an important role in medical information technology projects, as the implementation of these systems requires an understanding of the processes involved in them [14].

A RIS as defined by [15] is a computer system designed to support operational workflow and business analysis within a radiology department; it is a repository of patient data and reports which contributes to the electronic patient record (EPR) or electronic health record (EHR). [15] described the RIS as an imaging information system since it supports many additional specialists in areas including nuclear medicine, radiotherapy, and endoscopy.

As a RIS contributes to EHRs, then any errors in these systems propagate to affect the EHRs, which may put clinicians in a situation where they make wrong diagnosis and consequently put patients' lives at risk.

The interviews showed that one of the main concerns about adopting the RIS/PACS systems is the potential lack of reliability and thus lack of safety of these systems; this is due to the difference between the theoretical workflow and the operational workflow. Furthermore, even the theoretical workflow possibly has many problems with its safety, as where the safety issue was not addressed specifically during the workflow design. This leads to output failures of different parts of the workflow and eventually failure of the final output of the system. These failures of outputs can be defined by output deviations, where an output deviation outlines a set of Boolean expressions that shows the causes of the output failure, and the relationships between them. These causes can involve internal failures, input deviations, or both.

There is a scarcity of published literature addressing the problem of analysing the operational workflow and its safety. It is uncommon to have a formal automated safety analysis in healthcare for the management of healthcare systems' operational workflows such as the workflow within the radiology department. Little information is available regarding operational errors in RIS/PACS workflows (e.g. [7]). Research to date has not identified efficient automated approaches for workflow errors risk reduction. Many aspects of RIS/PACS design can be changed through the safety analysis of the workflow, as a flawed workflow design has the potential to decrease the efficiency and increase user errors during the operational phase of the workflow.

In the face of these limitations, this paper identifies potential significant errors in a RIS/PACS workflow by means of the following:

- Using an integrated safety analysis approach to analyse the safety of the RIS/PACS workflow

- Using the results of the empirical study to document and model both the detailed processes and the in depth tasks of one failure scenario of this workflow.

- Collecting data regarding occurrence of the workflow errors and their prevention in the same scenario environment.

- Discussing current approaches to reduce the risk of errors in the RIS/PACS workflow.

The following sections show the application of each step in the approach to analyse the safety of the workflow with this radiology department.

A. Requirements Generation

While documenting the requirements we found that the ideal architecture for a RIS has a hospital information system (HIS) which works as a master patient index, where data goes immediately to the RIS without the need for a technologist to enter any data.

In our case, the hospital combined the RIS and PACS and has them as a stand-alone departmental radiology system. They have a non-complete HIS that does not have full functionality

and is not connected to the RIS. All the data needs to be entered in the RIS by the clinicians. The information to be entered includes the following: Patient name, Patient National Number, Date of Birth (DoB), Age, Address, Patient medical Information, and Order Information.

After the above information is entered into the RIS either by the clinician (as in our case) or by coming immediately from the HIS, then this information (which includes patient's medical, administrative, demographics, and billing information) is kept in the RIS, in addition to the information which is added at the RIS to identify the examination order. These may include the following: Order ID, Order Description, Scheduling, Patient Arrival Information, and Examination Room Scheduling. This discussion considers the case where the clinician enters part of the information into the RIS, and there is some information that is entered into the RIS by another party who might be a radiologist. After that, the output of the RIS goes to the modality work list (MWL) where the orders are scheduled to be sent to the image acquisition modality. Here at the image acquiCsition modality, there is no chance for human error as the data comes immediately from the RIS. However, this database, which has all the scheduling information and orders information, is open to hardware and software errors. At the image acquisition modality the patient is supposed to have the examination that is specified in the order. The output of the image acquisition modality is patient id, patient name and the image itself.

After that, these outputs are sent automatically to the PACS which archives them and then sends them to the diagnostic workstation to be seen by the radiologist. The radiologist is now able to interpret examinations from several clinical sites and/or hospitals (in the case of Tele-radiology), and produce a report as an output. This report is to be passed to the clinician to make the diagnoses and give a medicine or recommend for another procedure such as an operation.

This paper considers one of the workflow scenarios; the purpose is to analyse possible failures of this scenario and to find out the root causes of these failures. This scenario is the workflow for a computerised tomography (CT) scanner. A CT scanner creates cross-sectional images of the body using X-rays; the result is a very detailed 3D view of the body interior. CT scans are used to make a cancer diagnosis or assess the effects of cancer treatment.

When the patient sees the clinician, the clinician decides if there is a need for a CT scan. Once a CT scan is recommended, the risk of exposure to radiation is considered before deciding to send the patient to the exam. This is because the accumulative amount of radiation the patient is exposed to has a potential risk for the patient, so clinicians recommend it when they think that the benefits will exceed possible risks. In order to consider the amount of radiation, in most cases the date of the last CT scan must be considered by the clinician before such a decision can be confirmed. Moreover, a pregnancy check must be done to make sure that the woman who will start the exam is not pregnant.

Commonly, patients who will receive a CT scan must follow certain preparation guidelines. These include no eating for two hours before the appointment, and drinking 500 ml of water over this time. The water is useful to hydrate the patient before having the Contrast Media (CM) for the CT scan. Another preparation guideline is to ask the patient to drink another 500ml of water after arriving to the waiting area. It also helps to show the bladder on the scan.

Verbal verification by the radiologist is needed to check these preparations with the patient together with other preparations such as ensuring there is no metal present (e.g. wearing of a metal belt, or jewels or having an internal device inside their bodies). Moreover, verbal verification of the patient's DoB at this point plays an important role in correcting any previous errors in the DoB, as the DoB is important in determining the amount of CM and the amount of radiation. Some patients may require a blood test before CM can be given.

An injection of the contrast is often given before or throughout the scan. CM contains iodine and appears as white areas on the scans, which help the radiologist to differentiate between certain organs or tissues and the other structures. The contrast may be ingested as a drink, or injected around the required area, or given via a cannula which is placed in the patient's arm prior to the scan. Again, verbal verification is required here to confirm any allergies and medications that the patient takes in order to judge the suitability of the injection and to minimise interactions with other medications.

Typically, people who feel claustrophobic do not have problems with CT scan as they might have with other scans, like Magnetic resonance imaging (MRI). However, the radiographer should check this with the patient before the scan, as if the patient thinks that he is expecting to feel this way then an injection may be given before the scan to calm the patient.

After the scan is finished, the patient should be asked to wait for an hour at least after the injection to make sure the patient is in good health, and he/she did not have allergic reaction to the CM injection, because people sometimes have different reactions; in these circumstances, medical staff should be able to manage different reactions appropriately. The radiologist then should give some instructions to the patient to follow once he goes home, for example, again asking the patient to drink 500ml of water to rehydrate the body after the CM injection.

B. Workflow Modelling

Workflow model should specify the systems involved in producing a medical service and different agents who are interacting with these systems. Moreover, it should specify the dataflow as well as the sequence of event. The following figure shows the workflow within the considered radiology department. Matlab Simulink was used for the modelling process. The information from the EHR is relayed back to the HIS component.
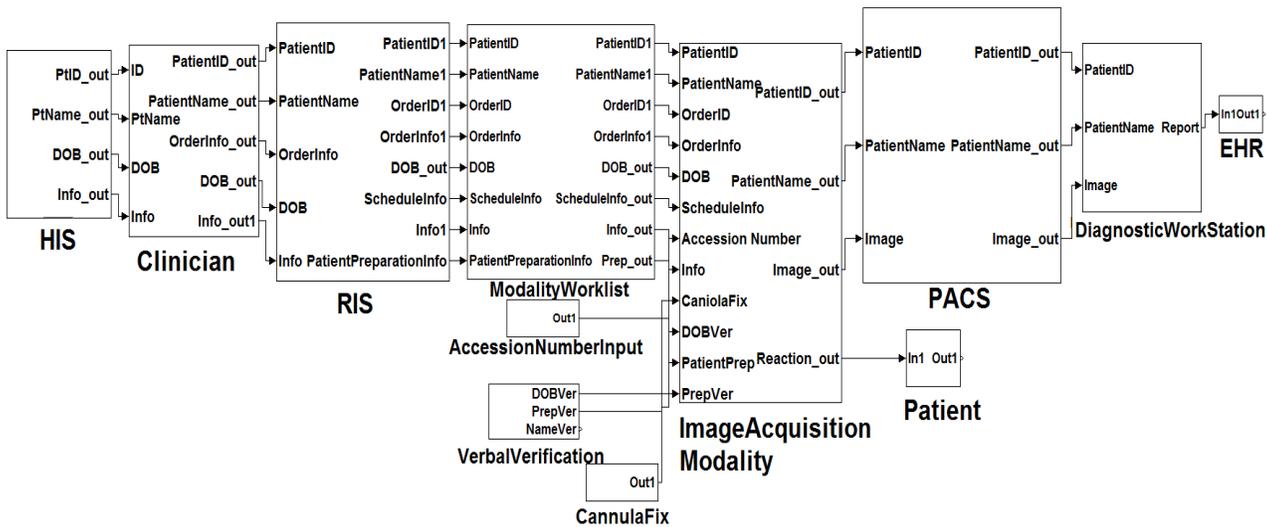
Fig.2.   RIS/PACS Workflow

The EHR is modelled as a subcomponent of the HIS and it has the following information:
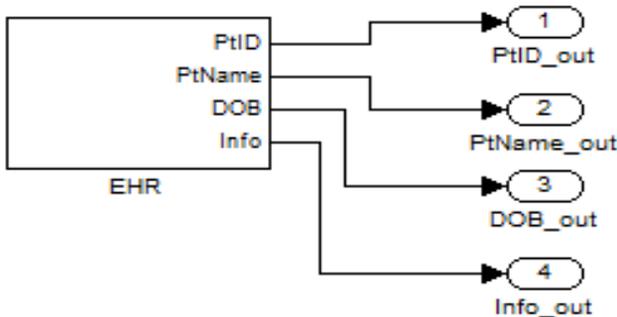
Fig.3.   EHR Component

The CM is modelled as a separate component (subcomponent of the ImageAcquisitionModality):
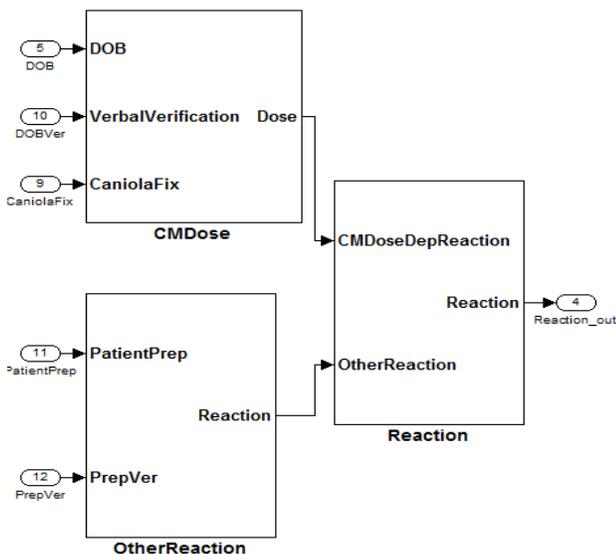
Fig.4.   The CM Component

C. Hazards Identification

After the analyst builds the model, the accuracy of the model is discussed with the healthcare experts in the hospital as well as the possible ways the whole workflow can fail. Possible faults in each component are specified. After that the analyst job will be to prepare the logical failure expressions which are appropriate for the failure annotation.

Errors may happen at any point where there is a data entry. This paper focuses on the failures caused by DoB errors in the CT scan workflow scenario.

A CT scan is considered as a safe procedure, although there can be reactions to contrast media CM which usually cannot be predicted [16].

For example, the dose of contrast media which is given to the patients is different for adults and children. Therefore, date of birth is an important factor for specifying the amount of CM to administer. Giving the patient an overdose of CM has reactions that affect patient health and put the patient in a hazardous situation.

Faults may occur at different points in the workflow and need to be identified.

D. Failure Annotation

Failure annotation is performed using the HiP-HOPS tool; all the components need to be annotated with possible faults. HiP-HOPS then analyses the model to give the fault trees that detects the possible failures and provide the root causes for them. In addition, it provides us with information about their effects on the output of the workflow.

HiP-HOPS has a simple language for annotating the components with reusable failure logic. For example,

O-Out = O-In or InternalFailure

On the left is the output deviation, which represents a failure propagated from an output port of the component. On the right is the cause of that deviation, consisting of the basic deviations or basic events. Both input and output deviations

consist of a user-defined failure class representing the type of failure (e.g. O= omission) and the name of the port question. We can annotate the same component with multiple output deviations, and failure annotations as well can be applied to subcomponents, in other words, it can be hierarchical.

So, the interview's data are analysed to document the RIS/PACS system's workflow. Then the documented information is used to model the workflow to enable the automated analysis. After that, possible hazards are identified and failure expressions are now ready to be annotated into the model for several scenarios. As explained earlier this failure expressions describes how a failure in the component output is caused by a propagation of failure from the component input or the internal malfunction of the component itself. Failure is represented in the format of "FailureType-ComponentName.ComponentPort" in HiP-HOPS.

The first scenario analysis focuses on having side effects or bad reactions by the patient. As described by the system architecture, the effect on the patient is considered as an 'output' component. This failure is represented by the value failure of the patient component, and so is referred to as V-Patient.Out1.

The patient's DoB is entered into HIS together with other information. Value failure of DoB which could be caused by wrong data entry is represented as V-DoB_out, also omission of the DoB causes problems and it is classified here as output deviation of the HIS. Omission of DoB is represented here as O-DoB_out. Moreover, HIS internal malfunctions can cause the output failures of the HIS; these are represented as HWError, SWError, and DataEntryError.

Similarly, the clinician — who is included in the workflow as a separate component — can have output deviations. Clinician might make data entry errors which are represented here as IDDataEntryError or DoBDataEntryError. The output deviations are represented as V-PatientID_out and V-DoB_out.

RIS internal malfunctions may include software or hardware malfunction, represented as HWError, SWError. RIS as well as potentially receiving the wrong DoB from the Clinician, represented as DoBDataEntryError. In addition to these malfunctions, RIS may suffer from failure of the preparation data, which is PrpDataEntryError. Therefore, output deviations at RIS could be the omission of DoB or having the wrong DoB or having the wrong preparation information or omission of preparation information; these are represented respectively as: O-DoB_out, V-DoB_out, V-PatientPreparationInfo, and O-PatientPreparationInfo.

ModalityWorkList is a database, which keeps orders' scheduling information and patients' information. It can have two basic events, which are software error or hardware error. These are represented as SWError and HWError respectively. Each of the ModalityWorklist inputs has its own failure but in the first scenario, some failures have been considered and the others are ignored as they are assumed to be free from failures. The failures which are to be analysed are: the failure of the value of the DoB and the failure of the preparation information output either as a value failure or omission of this value. These are represented as V-DoB_out, V-Prep_out, and O-Prep_out.

When it comes to the image acquisition modality itself, at the time of the test the radiologist should verify some information with the patient, e.g. DoB, name, and preparations for the test. The process of verbal verification is represented as a separate component which may have two basic events, both human errors; they are represented as: DoBHumanError and PrepHumanError. Failures of the output of this component are represented as: O-DOBVer and O-PrepVer.

Fixing the cannula for the contrast medium is considered as well as a separate component, and annotated with the failures that might be a human error (represented as HumanError); the output failure of this component is represented as V-Out1.

The CM dose is considered as a subcomponent of the image acquisition modality and failure of this is giving the wrong dose for the patient. This is represented as V-Dose, which can be caused by either wrong dose calculation or wrong measurement. Other reactions are considered as well as subcomponents of the image acquisition modality component, which may have a failure that is represented as V-Reaction, where the patient has some reactions or Side effects when he is not supposed to have them. These kinds of reactions that happen according to not following the preparation guidelines by the patient are separated from the CM dose-dependent reactions.

The output of the CMDose component and OtherReactions component goes to the Reaction component. This separate component is annotated as well with possible failures. The output deviation of this component is having any type of reactions by the patient. This is represented as V-Reaction.

The reactions component is connected to the patient who is having these reactions. The image output is connected to the PACS component that receives the images and archives them into a database.

We did not annotate both the PACS and the diagnostic workstations component with failure information for the purpose of this scenario. We assumed that they only propagate failures. A comprehensive analysis must consider failures of these components and annotate them with all possible errors to get the root causes for the other possible failures of the workflow.

There are other scenarios that may possibly cause defective results, but again, for simplicity, they are not covered in this paper. For example, when the patient gives information to the clinician, the patient might not tell the right information about his situation and the clinician might not check. Those two conditions together result in creating the wrong history for the patient. When the clinician has the wrong information, he or she will ask for the wrong exam order that in turn causes the wrong examination description. At the time of the examination, if the patient did not tell and the radiographer did not verify this, and he or she has the wrong exam description, these conditions together might give a false report for the patient, which results in an incorrect procedure or the wrong medication.

Another failure that can potentially cause patient harm but is not considered in this paper is when images are mislabelled for the wrong patient and/or the wrong study. These kinds of

failures result in images that are incorrectly associated with the patient's EHR and may lead to incorrect diagnoses, medication, or procedures.

Other failures might happen because of an incorrect entry for the DoB of the patient, which occurs when the clinician enters the wrong DoB in both the HIS and the RIS. These faults together result in the wrong DoB of the patient which cause an incorrect dose of both radiation and the CM. Here the patient is under the risk of extra dose of radiation and dose dependent reactions of CM. The dose dependent reactions of CM are analysed in this paper.

E. Fault Tree and FMEA Synthesis and Analysis

We annotated the components of the model with the corresponding logical failure information and then performed the root cause analysis. At present, as sufficient numerical data for the components are not available therefore the numerical data associated with components are not entered as part of the annotation. HiP-HOPS synthesises and analyses the system fault trees and produces the FTA and FMEA results, which shows how the value failure in an input and the component failures (or their combinations) can lead to the failure in causing unintended reactions or side effects towards the patients.

The following figure shows the FTA result. For simplicity, V-Reaction is represented as Unintended Reaction in the FTA and FMEA table:
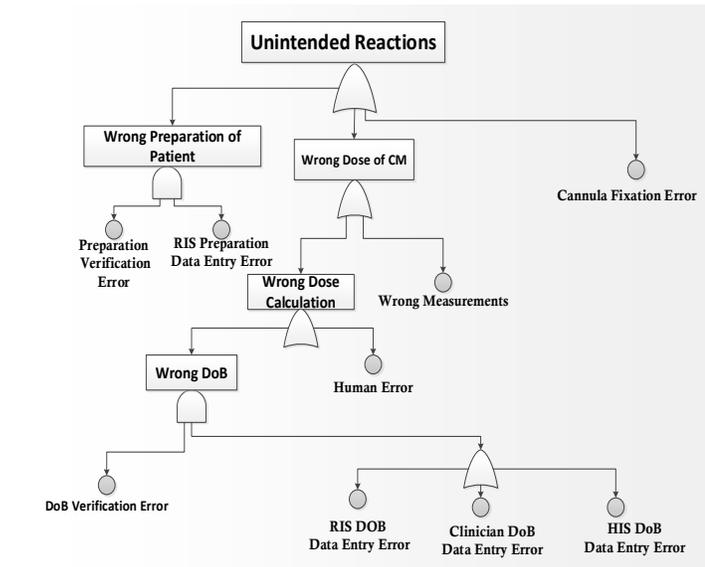
Fig.5.   FTA Result

The following list shows the MCS [4] from the FTA:

TABLE.I.   MCS FROM FTA

Cannula Fixation Error
Wrong Measurement
Calculation Human Error
Preparation Verification Error AND RIS Preparation Data Entry Error
DoB Verification Error AND RIS DoB Data Entry Error
DoB Verification Error AND Clinician DoB Data Entry Error
DoB Verification Error AND HIS DoB Data Entry Error

The following table shows the resulting FMEA table of the direct and further effects:

TABLE.II.   FMEA TABLE

Failure Mode	System Effect
Component: Cannula Fixation	
Human Error	Unintended Reactions
Component: Clinician	
DoB Data Entry Error	Unintended Reactions
Component: RIS	
DoB Data Entry Error	Unintended Reactions
Preparation Data Entry Error	Unintended Reactions
Component: Verbal Verification	
DoB Human Verification Error	Unintended Reactions
Preparation Human Verification Error	Unintended Reactions
Component: HIS	
DoB Data Entry Error	Unintended Reactions
Component: CM Dose	
Wrong Measurement	Unintended Reactions
Calculation Human Error	Unintended Reactions

To summarise, the FTA and FMEA results show that the following failures may lead to the failure of the first scenario (which is in this case getting unwanted reactions by the patient):

- Human error in fixing the cannula for the CM, where the radiologist or the nurse makes an error in placing the cannula prior the scan. This mistake cause problems for the patient as the CM is injected through the scan, which might lead to both side effects of the CM or extra dose of radiation because radiologist might need to repeat the scan.

- Data entry error for the DoB by the clinician combined with an error in the verbal verification of the DoB by the radiologist at the time of the scanning. This combination of errors might lead to an extra dose of radiation and/or extra dose of CM, which may put patient's life at risk.

- Data entry error for the DoB by the radiologist combined with an error in the verbal verification of the DoB by the radiologist at the time of the scanning. Again this focuses our attention on the importance of the verification of the DoB by the radiologist at the time of the scanning.

- Data entry error for the preparation guidelines by the radiologist combined with an error in the verbal verification of the preparation guidelines by the radiologist at the time of the scanning. This means, if the patient received the wrong preparation guidelines or did not receive them at all, then at the time of the scanning, if the radiologist does not make sure about their accuracy (and whether they were followed by the patient or not), the patient will experience the reactions.

- Data entry error for the DoB in the HIS combined with an error in the verbal verification of the DoB by the radiologist at the time of the scanning.

- Wrong measurements to calculate the dose of CM can directly cause the unwanted reactions. This might happen because of not understanding the units of measurements, or using wrong equipment to measure the dosages.

- Human error in calculating the dose can directly cause the reactions. This may happen through making slips in calculations that result in wrong dose.

This means that if there is any error in the data entry in HIS, clinician, and the RIS, combined with a situation where the radiologist does not verify (or verifies incorrectly) the data for DoB or preparation information, the unintended reactions towards the patient will occur.

These errors can be avoided by adding extra functionality to the HIS or RIS or both of them (for example, bar coded patients help to avoid data entry errors by radiologists and clinicians). Moreover, adding extra tasks in the workflow may help to avoid the errors.

Human error in fixing the cannula for the CM also contributes directly to the unintended reactions. So, radiologists or nurses who perform this task should be informed about potential failures which it may cause and about their direct and indirect effects on the patient. As mentioned earlier, the numerical data for the components of the system in the case study are not considered; hence, quantitative analysis is not performed in this paper.

F. New Workflow Design and Accreditation of the Workflow

A new workflow can be proposed with some additional tasks and avoiding the potential failures. Existing critical tasks may be highlighted as important tasks to focus the attention of the healthcare team to the importance of this task.

The model optimisation capability of HiP-HOPS [17] can also produce different alternative models to help to achieve safety requirements, and in doing so it can assist in selecting component and subsystems among different alternatives as well as helping to decide the level and location of replicated components.

G. Evaluation of the Approach

To evaluate our approach the concerns about the application of the approach to analyse the safety of clinical workflows were discussed with both healthcare and technical experts in a systematic manner. The approach is found to have the required usability; it is tool based and a user friendly graphical modelling tool (Matlab Simulink) is used for the modelling process. Of course, the analyst's ability to create the models has a non-trivial impact on the accuracy of the models. However, the approach suggests that models should be checked and approved by healthcare experts before proceeding with the other steps.

In comparison with other methods which are already in use for safety analysis (e.g. Bayesian Networks (BNs) which is primarily used for quantitative analysis), the modelling phase in our approach requires less technical knowledge. For example, modelling workflows in BNs to facilitate safety analysis requires specialised technical knowledge and there are no well-defined rules to create BNs of a workflow for safety analysis purposes. As a result a system can have a number of equivalent BNs and if the causal relationships between different nodes are not well-defined then BNs can become unnecessarily complex and non-coherent. Therefore BNs of a workflow may be not be understandable by healthcare experts, thus it may not be possible to involve the healthcare experts in the early design phase though their involvement is highly required. However, BNs are efficient methods to perform quantitative analysis under uncertainty and a coherent and relatively simple BN can be created by translating other representations, e.g. Fault Trees into BNs. As mentioned earlier, analysts can benefit from the capability of BNs by using them in the stages where the healthcare experts are not involved. In our approach, healthcare experts are no more involved in the safety analysis process after the workflow has modelled. Therefore when fault trees are generated by the HiP-HOPs tool then they can be translated into BNs, thus benefit from the strength of the BNs while involvements of the healthcare experts are also ensured.

Even though other modelling techniques can be beneficial, for example, finite state machines can be used to model the workflow, it is not easy to analyse the state machines directly to obtain safety related information about the workflow. In this case, state machine based model will be required to be transformed to other models e.g. fault trees, Markov chains, Bayesian Networks. Another issue of state machine based approach is that they increasingly face state explosion problem, i.e., for a relatively complex workflow, number of states required to model the complete failure behaviour of the workflow grows exponentially with the number of components of the workflow, thus are difficult to create and analyse.

The process of hazards identification is done by the analyst in cooperation with the healthcare experts. The hazards should be specified for each component and possible failures for each component and their causes are discussed. After that the process of failure annotation needs to be done by the analyst who should have experience in using the HiP-HOPS. In terms of usability, HiP-HOPS has a graphical user interface which is easy to use, and it does not need an expert as it is uncomplicated in comparison with the other methods.

In comparison with other model checking [18] or simulation approaches, HiP-HOPS is less automated than these approaches. However, it is generally faster and more scalable and can be used to complement other techniques such as simulation. Recent work on the systematic application of HiP-HOPS and model checking [19] also opens the opportunity to extend the analysis with model checking in future. HiP-HOPS also serves as a useful foundation for related technologies such as optimisation [20]. And, the model optimisation capability of HiP-HOPS can produce different alternative models to achieve safety requirements and can help to select component and subsystem among different alternatives as well as helps to decide the level and location of replicated components.

Workflow models are reusable and maintainable; that is if a certain workflow has been done for a certain department, then the analyst can use it as a subcomponent in another workflow and it can be easily maintained as well if there is a need to do this.

The safety analysis approach has the potential to affect both the workflow and the clinical service quality. The approach supports a large part of the workflow development process, in particular the design phase of the development process. The clinical service quality (which is the output of the clinical workflow) is improved and maintained through specifying the exact safe steps or baths which can lead to the service.

In the case where a workflow management system is required, then the approach has the potential to help in developing a reliable workflow management system; as the approach improves the quality of the design phase which leads to a better quality in the following software engineering stages.

It is generally accepted that a high quality product requires a high quality design. Theoretically, we can generate a hypothesis that if a high-quality safety analysis approach is maintained, its output has the potential to help in preparing high-quality and reliable clinical workflows.

Data on the applicability of the proposed approach were gathered through an informal testing shown positive usability and effective results. Our results were discussed with experts in the hospital where the data were collected and they appreciated the ability of the approach to focus on processes and how this could be employed for several applications in clinical workflows. Moreover, having the fact that our analysis results are happening in the hospital as actual failures has the potential to validate our approach.

Our approach drew the map for the root causes of these failures. This is the major contribution of this work as to date there is a lack of automated tools which allow the modelling and analysis of real-world workflows. The approach provides an effective means to accomplish this goal, is able to provide a valid theoretical framework consisting of modelling the processes and sub processes and their error analysis. The study findings contribute towards a larger research effort being proposed for reducing medical errors and enhancing patient safety.

Dependability can be improved based on the analysis from the tool: the workflow can be adapted, with workflow components substituted with more reliable components, components can be replicated to introduce redundancy and the frequency of maintenance can be increased for critical components.

IV. CONCLUSIONS AND CONTRIBUTIONS

The automated identification of these root causes allows greater understanding of the factors contributing to the undesired event which can potentially lead to a serious clinical risk. This enables the identification of weak points, which could then be effectively addressed and improved.

The simple act of undertaking a safety analysis in this way helps to improve understanding of the behavior of the workflow and its potential for failure, thus highlighting areas where additional checks or amendments to the workflow need to be introduced. The automation then additionally helps deal with the complexity and time cost issues, offering benefits over a simple manual analysis. While in this case there were only order 2 MCS, more comprehensive analyses might introduce even higher order MCS that are even more difficult to spot manually, potentially highlighting issues that are not even apparent from a manual analysis.

For example, through the simple structure in this example, the application of HiP-HOPS shows the ability to systematically assist in the identification of failures in the workflow (i.e. failure in the verbal verification or failure in the data entry of the DoB) and the identification of the failures in the system (i.e. hardware or software error in the MWL). This information can be used to guide the improvement in the design of both the system and the workflow. The system can be improved by targeting the areas where highly-reliable components and fault tolerant mechanisms can be prioritised and introduced to make the architecture more robust and fault tolerant.

Moreover, the workflow can be improved by designing the workflow in a way which takes the safety analysis into consideration and to use the results of the analysis to target

areas where reliable components (in this case the components are processes and tasks) can be introduced. The workflow should have an exact determination of the processes, tasks, and the procedures which must be done by each party.

Having this detailed workflow with a detailed analysis of the failure behaviour can enable healthcare organisations to develop material to be used by medical staff in safety training workshops. These workshops should help the medical staff to build safety awareness that may be useful to avoid the expected failures in the workflow.

Using HiP-HOPS in workflow analysis in general has the potential to give effective analysis by detecting possible design flaws early before serious problems happen. This also helps to provide the medical staff the awareness they require and aids in redesigning the workflow to produce an effective and fault free workflow.

Moreover, such modelling of the workflow and the analysis results can also be used as an educational tool for training of radiologists, nurses, and clinicians. This helps the trainees in identifying errors and preventing the potential errors from leading to adverse events.

The example presented in this paper is based on one scenario, while different scenarios need to be modelled and analysed to get a comprehensive analysis of the workflow. Moreover, conducting research of this nature on only one location is limiting, and having more sites opens a wider range of failures determination.

V. FUTURE WORK

HiP-HOPS is designed to consider local failures, where each component and its outputs/inputs has its own failure data. In clinical workflows, sometimes we may have the case where all components share the same cause of failure (e.g. human failure). The common cause failure idea allows this to be modelled, and this could be an improved way to model human failures in future.

REFERENCES

[1] L. Kohn, J. Corrigan, and M. Donaldson: 'To err is human: building a safer health system' (National Academies Press, 2000).

[2] M. Ruffolo, M. Manna, V. Cozza, and R. Ursino: 'Semantic clinical process management', in CBMS, pp. 518-523

[3] L. Al-Qora'n, N.Gordon, S. Sharvia, M. Walker, Y. Papadopoulos: 'An Approach to Safety Analysis of Clinical Workflows', Athens: ATINER's Conference Paper Series, No: COM2014-1157.

[4] W. Vesely, F. Goldberg, N. Roberts, and D. Haasl: 'Fault tree handbook', No.NUREG-0492, Nuclear regulatory comission washington dc, 1981.

[5] L. Ward, M. Lyons, S. Barclay, J. Anderson, P. Buckle, and P. Clarkson: 'Using fault tree analysis (FTA) in healthcare: a case study of repeat prescribing in primary care', Patient Safety Research: Shaping the European Agenda, 2007.

[6] E. Ekaette, R. Lee, D. Cooke, S. Iftody, and P. Craighead: 'Probabilistic fault tree analysis of a radiation treatment system', Risk analysis, 2007, 27, (6), pp. 1395-1410.

[7] H. Abujudeh, and R. Kaewlai: 'Radiology Failure Mode and Effect Analysis: What Is It? 1', Radiology, 2009, 252, (2), pp. 544-550.

[8] Y. Papadopoulos, and J. McDermid: 'Hierarchically performed hazard origin and propagation studies': 'Computer safety, reliability and security' (Springer, 1999), pp. 139-152

[9] J. Esary, and F. Proschan: 'Coherent structures of non-identical components', Technometrics, 1963, 5, (2), pp. 191-209

[10] A. Bobbio, L. Portinale, M. Minichino, and E. Ciancamerla: 'Improving the analysis of dependable systems by mapping fault trees into Bayesian networks', Reliability Engineering & System Safety, 2001 b, 71, (3), pp. 249-260.

[11] L. Al-Qora'n, S. Sharvia, N. Gordon, and Y. Papadopoulos: 'Safety Analysis of a Remote Patient Monitoring System with a Guideline Based Decision Support', Luxembourg: Porceedings of Med-e-Tel, April 2013 pp. 322-327

[12] S. Boochever: 'HIS/RIS/PACS integration: getting to the gold standard', Radiol Manage, 2004, 26, (3), pp. 16-24

[13] [13] G. Paré, D. Aubry, L. Lepanto, and C. Sicotte: 'Evaluating PACS success: a multidimensional model', Proceedings of the 38th Annual Hawaii International Conference on. IEEE, 2005, pp. 147c-147c

[14] A. Ouvry: 'Workflow analysis and modeling in medical IT projects', Medicamundi, 2002, 46, (2), pp. 47-55

[15] The Royal College of Radiologists: 'Radiology Information Systems', 2008,[Online].Available: http://www.rcr.ac.uk/docs/radiology/pdf/IT_guidance_RISApr08.pdf

[16] F. Stacul: 'Managing the risk associated with use of contrast media for computed tomography', European journal of radiology, 2007, 62, pp. 33-37

[17] Y. Papadopoulos, M. Walker, D. Parker, E. Rüde, R. Hamann, A. Uhlig, U. Grätz, and R. Lien: 'Engineering failure analysis and design optimisation with HiP-HOPS', Engineering Failure Analysis, 2011, 18, (2), pp. 590-608.

[18] E. Clarke, O. Grumberg, and D. Peled: 'Model checking' (MIT press, 1999).

[19] S. Sharvia, and Y. Papadopoulos: 'an Approach towards Integrated Safety Assessment ', in Proceedings of the IEEE 7thInternational Conference on Automation Science and Engineering. Trieste, August 2011.

[20] M. Adachi, Y. Papadopoulos, S. Sharvia, D. Parker, and T. Tohdo: 'An approach to optimization of fault tolerant architectures using HiP-HOPS', Software: Practice and Experience, 2011, 41, (11), pp. 1303-1327

Benchmarking the Higher Education Institutions in Egypt using Composite Index Model

Mohamed Rashad M El-Hefnawy
Computer Engineering Department
Arab Academy for Science &
Technology Cairo, Egypt

Ali Hamed El-Bastawissy
Computer Science Department
Modern Sciences and Arts
University Cairo, Egypt

Mona Ahmed Kadry
Computer Science Department
Arab Academy for Science &
Technology Cairo, Egypt

Abstract—**Egypt has the largest and most significant higher education system in the Middle East and North Africa but it had been continuously facing serious and accumulated challenges. The Higher Education Institutions in Egypt are undergoing important changes involving the development of performance, they are implementing strategies to enhance the overall performance of their universities using ICT, but still the gap between what is existing and what is supposed to be for the self-regulation and improvement processes is not entirely clear to face these challenges. The using of strategic comparative analysis model and tools to evaluate the current and future states will affect the overall performance of universities and shape new paradigms in development of Higher Education System (HES), several studies have investigated the evaluation of universities through the development and use of ranking and benchmark systems**

In this paper, we provide a model to construct unified Composite Index (CI) based on a set of SMART indictors emulate the nature of higher education systems in Egypt. The outcomes of the proposed model aim to measure overall performance of universities and provide unified benchmarking method in this context. The model was discussed from theoretical and technical perspectives. Meanwhile, the research study was conducted with 40 professors from 19 renowned universities in Egypt as education domain experts.

Keywords—Key Performance Indicators; Composite Index; Analytic Hierarchy Process; Performance Measurement; Higher Education Institutions; Ranking Systems; Benchmark Models

I. INTRODUCTION

Egypt has the largest education system in the Middle East and North Africa and it has grown rapidly since the early nineties, according to the Human Development Index (HDI) 2013 report, Egypt is ranked 121 in the HDI, and number 7 in the Medium HDI countries in Africa [1]. The higher education system in Egypt has around 2,646 million students were enrolled across all tertiary levels. They attended one of 24 public universities, including Al-Azhar University, the oldest continuously running university in the world; 18 private universities including non-for-profit Nile University; 3 private Academies; 58 public non-university institutions and 137 private higher or middle institutes, as shown in Table I. Non-university institutions are middle technical institutes offering two-year courses and higher technical institutes offering four-year courses. Since 25 January revolution, seven new universities have been created. Based on strategic plan of Supreme Council of Universities (SCU) in Egypt [2], about

4,771 million students are expected to enroll in 2021 across all tertiary levels.

TABLE I. FACTS ABOUT HIGHER EDUCATION SYSTEM IN EGYPT

Number of HEI (in academic year 2012/2013)				
240				
Public	Private	Private Academies	Public non-university institutions	Private higher or middle institutes
24	18	3	58	137
Number of students (in academic year 2012/2013)				
~ 2,646 million students				

The Ministry of Higher Education supervises the tertiary level of education and in order to improve the quality of a university services, it has established a committee for the reform of higher education known as the Higher Education Enhancement Project (HEEP) committee. Higher education reform strategy included 25 projects addressing all the reform domains up to our knowledge till 2017, the most important project know as Information and Communications Technology Project (ICTP). This project adopted a service category stakeholder approach to quantify the ICT needs of academic institutes [3]. The services include IT infrastructure improvement and provide integrated network infrastructure between all universities; implementing unified Management Information System (MIS) [4] and Decision Support System (DSS) as administrative computing systems for each university and technical institute; implementing a web portal for each university; establishing a union catalogue index for the university libraries; training employees and staff members on IT usage; implementing unified eLearning solution and providing access to digital libraries and federated search tools to Egyptian scholars . Each higher education institute has all of the above services and resources, which has a well overall impact on performance of the higher education institute. Otherwise, no single university, however large, can encompass all knowledge. Every university has to make choices. It is demanding to be world-class university [5] in even a few academic area. Each university has to priorities the use of its resources and use them to improve overall performance and increase service quality. Knowing whether it is succeeding in its aims is another more demanding level of difficulty. The key challenging question is how universities' leaders will

know where their institutions stand and how they can be improved [6, 7]. The gap between what is existing and what is supposed to be for the self-evaluation and improvement processes is not entirely clear. For that, the comparative experiences from other universities may help shed the light on possible solutions to provide both strengths and weaknesses in the present strategic context.

In the literature, the methodology of university ranking has gained a considerable importance among higher education institutions [8, 12], a fair view can usually be drawn from the outcomes of indicators [9]. Common stakeholders in higher education, decision makers and researchers are definitely keen to know the position of the institutions and figure out the key reasons behind ranking results and outcomes to improve strategic planning, reviewing overall performance, improve operations, change management, compare performance with comparable institutions or with "best practices" benchmarks and to assist their institutions in evaluation, decision making, and improvement processes [10, 23].

This paper provides theoretical and technical framework as a benchmark for measuring the overall performance of higher education institutions in Egypt .The proposed model creates a mixed criteria Composite Index (CI) relative to a set of weighted and aggregated Key Performance Indicators (KPI), illustrated in a scientific manner. Attention is focused on using Analytic Hierarchy Process (AHP) to weight and aggregate [29, 11] the indicators used by the proposed composite index.

This paper is structured as follows: Section 2 describes the related work. Section 3 describes the used analysis method .The use of the method is illustrated by a case study in Section 4. Section 5 demonstrates experimental results. Section 6 presents conclusions and suggestions for future work.

II.   RELATED WORK

A. Ranking Systems

There are about 50 ranking systems in use around the world which use several of different key performance indicators. Several academic institutions, media organizations, governmental or non- governmental agencies [10] have already conducted ranking methodology on global, regional and national bases for higher education systems. The countries, which conduct rankings of universities, departments or programs, include Australia, Canada, China, Hong Kong, India, Spain, UK, US, etc. These countries have diverse systems of ranking based on different criteria covering indicators related to students, faculty, research publications, research funding and grants, alumni donations, graduation rates, social mobility, ethics of service to country, peers, finances, infrastructure, and reputation. These indicators broadly fall into seven subjects, student outcomes, financial inputs and outcomes, and staff outcomes, student experiences, learning and teaching outcomes, research outcomes and reputation [12]. The most famous ranking systems, Shanghai Academic ranking of world Universities (ARWU); Times Higher Education world university ranking (THE-QS); Webometrics ranking of world Universities; Taiwan Higher Education Accreditation and Evaluation Council (HEEACT).

The scope of world ranking systems are generic and do not focus on issues on issues related to region or local strategies, which may lead the decision maker to reform strategies based on ranking improvement for other countries rather than to do the right for the local setting. In general, the presence of Egyptian universities in global ranking systems is not bad, but this does not mean that Egyptian universities use these rates to improve their performance compared with others standard and leaders universities. The focus on the ranking systems results dispatch resources utilization away from success improvement factors related to certain issues because there are no criteria that measure those. In this paper, we focus on set of KPIs that are related to Egypt country strategy for higher education system.

B. Analytic Hierarchy Process (AHP)

As data analysis method, AHP was used in the study. AHP was firstly put forward in 1968s by Myers and Alpert and was developed and transformed to a usable model in resolutions of deciding problems as analytical techniques for complex decision making problems by Saaty in 1977s [13, 22]. In AHP method, quantitative values like price, weight, or area, or even subjective opinions or qualitative items such as feelings, preferences, or satisfaction, can be translated into measurable numeric relations. AHP belongs to the multi-criteria decision making methods (MCDM) group, an estimation method that is used in cases in which a decision hierarchy is present and that interpret per cent distributions of decisions points, in terms of factors affecting the decision. AHP on a decision hierarchy, is based on pairwise comparisons to determine significance values of decision points, in terms of the factors that affect the decision using a comparison scale identified beforehand. Significance differences, ultimately, turn into per cent distributions on decision points. AHP has been used in fields of higher education such as scientific analysis, ranking and evaluation systems [14, 15, 16, and 24].

Practice process of AHP consists of five steps. In creation of hierarchy, the first step, the purpose of AHP method usage is defined and the hierarchical structure related to the purpose is set forth (Wind and Saaty, 1980) and reflect the relationship between the purpose of the comparison and the result that is desired to be obtained. The hierarchical structure includes main and short-listed sub-criteria containing the decision goal. The second step is the arrangement of binary comparison matrixes in which main or sub-criteria are compared with alternatives between each other. While the mutual comparison of the factors are done, the evaluation scale recommended by Saaty (1994) defined in Table II is used. This scale consists of importance scales that are defined from1 to 9. A criterion which is compared with itself, is always assigned the value 1, so the main diagonal entries of the pairwise comparison matrix are all 1. The numbers 3, 5, 7, and 9 correspond to the verbal judgments 'moderate importance', 'strong importance', 'very strong importance', and 'absolute importance' (with 2, 4, 6, and 8 for compromise between the previous values). Every two factors are enumerated mutually with one of these numbers. While creating comparison matrices, it is questioned on how much important the factor on the row when compared the one on the column. The intermediate values are the values that can be chosen by the decision maker if he is in dilemma

between two main values. If criterion A, is of stronger importance than that of criterion B then the comparative value 7 is given, if they have equal importance then the comparative value 1 is given. In this case, the criterion B takes the value 1/7 or 1 when compared to A. Because the reverse of the same criteria are the reverse of the same point according to multiplication. The third step of AHP is the creation of normalized matrixes and the occurrence of each factor's significance level. The fourth step of the method is the determination of the fact that whether the matrixes are consistent or not; the fifth step and the last stage, on the other hand, is the assignation of priorities. These five steps belonging to AHP method were followed through, elaborated calculations and equations discussed in the following section.

TABLE II.　　AHP Scale For Pairwise Comparisons

The Fundamental Scale for Pairwise Comparisons		
Intensity of Importance	Definition	Explanation
1	Equal importance	Two elements contribute equally to the objective
3	Moderate importance	Experience and judgment moderately favor one element over another
5	Strong importance	Experience and judgment strongly favor one element over another
7	Very strong importance	One element is favored very strongly over another; its dominance is demonstrated in practice
9	Extreme importance	The evidence favoring one element over another is of the highest possible order of affirmation
Intensities of 2, 4, 6, and 8 can be used to express intermediate values. Intensities of 1.1, 1.2, 1.3, etc. can be used for elements that are very close in importance.		

C. Composite Indicators

Generally, a composite indicator is formed when individual indicators are compiled into a single index on the basis of an underlying model. It is tailored wide-model for benchmark simulation and comparative analysis purposes, comparisons can be used to illustrate complex and elusive issues in wide-ranging fields, e.g., environment, economy, society or technological development. Steps followed to construct any composite index are listed as follows, (1) theoretical and conceptual framework should be developed to provide the basis for the selection and combination of single indicators into a meaningful composite indicator under a fitness-for-purpose principle; (2) indicators should be selected on the basis of their analytical soundness, measurability, relevance to the phenomenon being measured and relationship to each other; (3) consideration should be given to different approaches for imputing missing values; (4) exploratory analysis should investigate the overall structure of the indicators, assess the suitability of the data set and explain the methodological choices like weighting, aggregation; (5) indicators should be normalized to render them comparable; (6) indicators should be aggregated and weighted according to the underlying theoretical framework; (7) analysis should be undertaken to assess the robustness of the composite indicator in terms; (8) testing should be applied in the real data; (9) attempts should be made to correlate the composite indicator with other published indicators, as well as to identify linkages through regressions; and (10) composite indicators can be visualized or presented in a number of different ways, which can influence their interpretation. Indeed, general public easily

can analyze composite index outcomes than identifying common trends across detailed indicators [17]. Bandura in 2006 considered an inventory of 165 composite indexes in the development space [18].

III. Analysis Method

In order not to sever the ties with the main theme of the study, findings related to demographic features, theoretical framework development, data selection, rehearse meetings and multivariate analysis steps with the ICTP professors as education experts have not been discussed here in detailed, also others points related to survey design methodology over data collection up to reliability and validity in practical testing are not discussed.

There are several techniques based on Multi-Criteria Decision Analysis (MCDA), we have selected the popular Analytic Hierarchy Process (AHP) this techniques support decision making through "Pairwise Comparisons" which uses comparison between pairs to build our weighted composite index. The block diagram in Fig.1 show the methodology used to collect responses from domain experts. The survey website populated the history database with records. Microsoft SQL Server stored procedures are invoked by an admin module to execute AHP algorithms then display the results to decision makers. In the algorithms, the priorities are calculated using the row geometric mean method (RGMM) [26], either aggregation of individual priorities (AIP) for individual participants, or an aggregation of individual judgments (AIJ) [25] based on the weighted geometric mean method (WGMM) of all participants' judgments can calculated, alerts and notifications for judgment exceed level of consistency are implemented. The consistency ratio (CR) and the geometric consistency index (GCI) are calculated. The judgment matrix, normalized principal Eigen vector and ranking of priorities are visualized in dashboard as presentation layer.
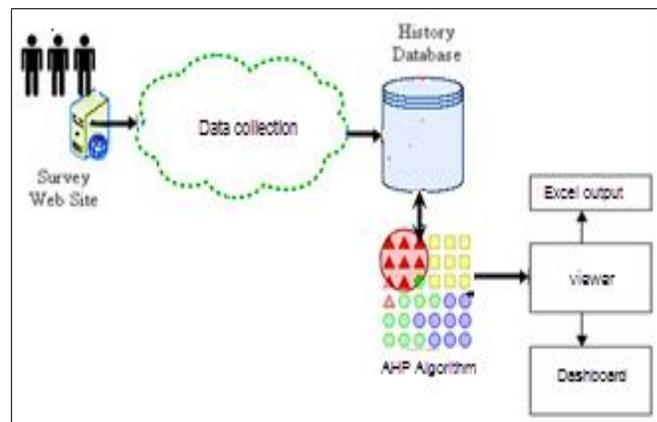
Fig. 1.　Block diagram of the proposed system

IV. Case Study

A. Constructing the Compsite Index

As discussed earlier, there are steps followed to construct the composite index, these steps has categorized and grouped into four phases as shown in Fig.2 phase one present the techniques used to develop the theoretical and conceptual

framework, using these topics (1) Theoretical framework; and (2) Data selection.
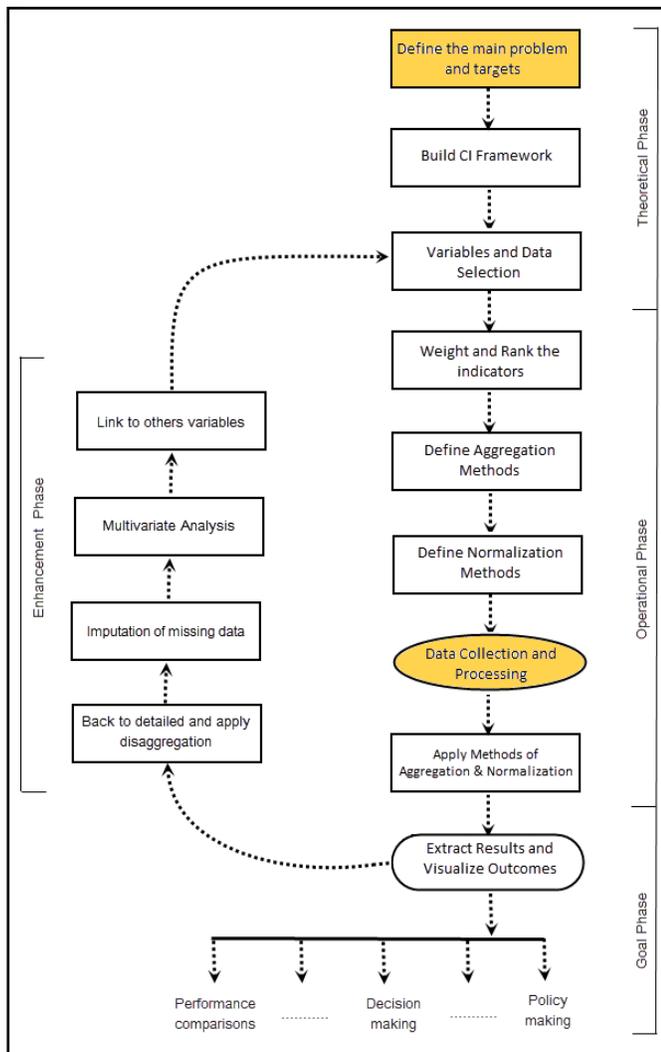
Fig. 2.   Life-cycle of proposed Composite Index

Fig. 3.   Hierarchic structure of the problem

Phase two present the operation and calculation methods used to build the proposed composite Index, based-on (3) Weighting and aggregation methods; (4) Normalization methods; and (5) Consistency and Uncertainty Analysis. Phase three focus on the results extraction, using (6) Presentation and Visualization techniques. Phase four focus on the mechanisms used to enhance the outcomes of composite Index after the dissemination stage, based-on (7) Back to details on the real data; (8) Links to other variables; (9) Imputation of missing data; and (10) Multivariate analysis; Meanwhile, the steps listed regarding phase four postponed as feature work.

B. The Theoretical Framework and Variables Selection

The first stage of case study implementation, hierarchy, has been generated and shown in Fig.3, which is basically composed by three levels: the goals, the main-criteria (KPIs' Areas), the sub-criteria (set of KPIs under each area) and the alternatives that used to achieve the research target.

The key performance indicators were discussed and ascertained by the education experts and these indicators were classified on the basis of business functions, each KPI has been carefully selected according to the global standard of higher education benchmark and belong to the Egyptian situation. Moreover, the KPIs follow the SMART criteria.

This means the measure has a specific purpose for the business, it is measurable to really get a value of the KPI, the defined norms have to be achievable, the improvement of a KPI has to be Relevant to the success of the organization, and finally it must be time bound, which means the value or outcomes are shown for a predefined and relevant period. The definitions of the indicators concerning the codes and main area illustrated in Table III.
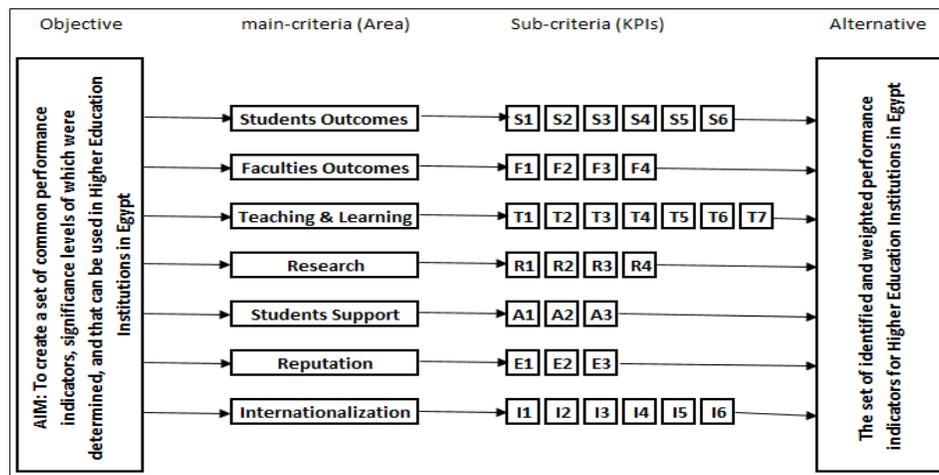
TABLE III.   HIERARCHICAL STRUCTURE OF THE PROBLEM

Area	Code	Indicator	Description
Students Outcomes	S1	Overall Student satisfaction (Survey)	Proportion of students expressing satisfaction with overall experience in student survey
	S2	Completion	Proportion of students scheduled to graduate succeeding in doing so (undergraduate - postgraduate)
	S3	Retention rate	Percentage of first-time freshmen students who return for the fall of the following year (dropout after one year & two year)
	S4	Granted Degrees	Number of degrees granted to students (Master-Doctoral)
	S5	participation	Percentage of students trained in public or private sectors annually
	S6	Scholarships and bursaries	Number of students receiving Honors Scholarship
Faculties Outcomes	F1	Faculty qualifications	proportion of academic faculty with earned doctorates to be equivalent to international proportions against number of junior staff
	F2	Attainment	Number of full-time faculty (teaching and administrative) with Offered degrees
	F3	Publications	Number of annual publications by faculty
	F4	Efficiency	Number of faculty engaged in international research activities
Teaching & Learning	T1	Faculty to student ratio	The ratio of full-time faculty to full-time students
	T2	Faculty to Administrative staff ratio	The ratio of full-time faculty to Administrative staff ratio
	T3	Satisfaction with teaching (Survey)	Proportion of students expressing satisfaction with teaching through student survey (Teaching Methodology , Staff , Courses)
	T4	Further study	Proportion of students pursuing further study within 12 months of graduation
	T5	Gender balance	The ratio of recognition of gender parity
	T6	Graduate employment rate	Percentage of alumni reporting being employed 24 months after graduation
	T7	Continuous improvement	Number of new programs offered by university
Research	R1	Citations per paper	Normalized average citations per paper
	R2	Papers per faculty	Average of research papers per faculty
	R3	Research excellence & outputs	Number of cited papers or articles in high impact journals
	R4	Research funding	Average Amount of postgraduate students from university budget of research
Students Support	A1	Environmental impact (Survey)	Performance against a Survey of environmental indicators
	A2	Low-income outreach	Proportion of students identified as low-income OR investment in promotion to low-income families
	A3	Internal migration	Proportion of overseas students from other governorates and non-foreign
Reputation	E1	International outlook	Number of ranking performance in specific disciplines or overall in one of international Ranking system
	E2	Distinguished Alumni	Number of Alumni whom listed in distinguished list
	E3	Prolific academic experts	Number of faculty members or Alumni achieving international recognition through awards (e.g. Nobel Laureates or Fields Medalists)
Internationalization	I1	International faculty	Proportion of international faculty
	I2	Visiting scholars	Number of visiting professors from international universities

I3	International students	Proportion of international students
I4	Outbound exchange students	Proportion of outbound exchange students
I5	Recognized accreditations	Number of Programs or Research projects accredited by national or internationally recognized and applicable standards
I6	Patents	Number of Active patents registered with national or international patent offices

Areas = main-criteria
Indicators = sub-criteria

C. Weighting and Scoring the Compsite Indectors

A web site was designed, developed and published to enable education experts to fill a pairwise comparison questionnaire, as shown in Fig.4, that illustrate the main aim of composite index.
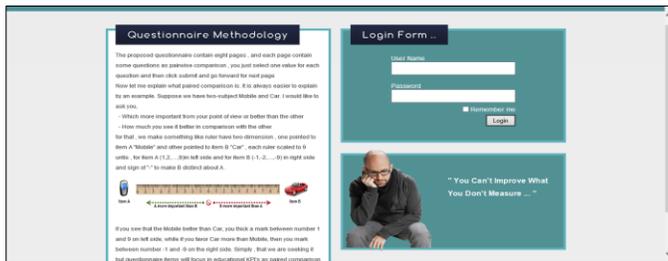
Fig. 4. The proposed website for data collection

The questionnaire has 90 pairwise comparison questions express hierarchical structure combination between main-criteria as main areas of KPIs , and sub-criteria as common set of performance indicators that used in higher education performance measurement, as shown earlier in Fig.3.

These indicators were classified on the basis of business functions to: (1) students indictors; (2) faculties indictors; (3) teaching & learning indictors; (4) research indictors; (5) students support indictors; (6) reputation indictors; and (7) internationalization indictors. The GUI shown in Fig.5 displays a part of pairwise comparison questions for a set of key performance indicators related to students support area.
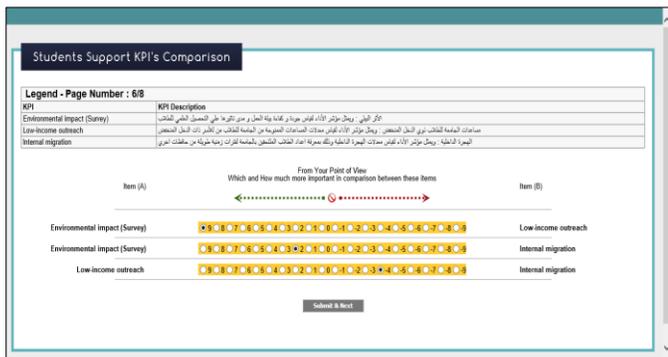
Fig. 5. Simple pairwise comparison page

Each question represented as two-dimensional ruler to identify importance of pairwise, as shown in Fig.6, importance selection based on evaluation scale defined in Table II.
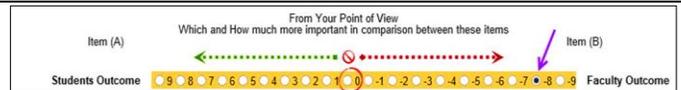
Fig. 6. Simple of question as two-dimensional ruler

One-to-one comparisons of factors correlatively, which is the second stage, the 1-9 evaluation-scale used is shown in Table II. The results of questionnaire survey given by the universities' professors in accordance with the evaluation scale are translated into pairwise comparison matrix then calculated in order to obtain geometric mean using the row geometric mean method (RGMM), and then to start weighting process.

In order to count the number of comparisons for combination of sixteen KPIs grouped in seven areas, can be represented as:

$$\frac{n(n-1)}{2} + \left\{ \sum_{i=1}^{n} m\,(m-1)/2 \right\} \qquad m, n > 0 \qquad (1)$$

Where n= number of main-criteria's items (KPIs Areas) and m= number of sub-criteria's (KPIs list that categorized under area).

$$= \frac{7(7-1)}{2} + \left\{ \frac{30 + 12 + 42 + 12 + 6 + 6 + 30}{2} \right\} = 90$$

This number reflects the total number of all pairwise comparisons rows and questionnaire questions, while the number of pairwise comparisons per each iteration for comparison matrix, can be represented as = $1+n = 1+7 = 8$. For each iteration, number of comparisons is $n(n-1)/2$.

The Aggregation of Individual Judgments (AIJ) [25] technique used in this case study. For the consolidated decision matrix Z that combines all k participants' inputs to get the aggregated group comparative matrix (judgment matrix), the weighted geometric mean method (WGMM) of the decision matrices attribute $r_{ij(k)}$ using the individual decision maker's weight w_k as given,

$$Zij = \exp \frac{\sum_{k=1}^{M} wk\,\ln rij(k)}{\sum_{k=1}^{M} wk} \qquad (2)$$

The pairwise comparison of attribute *i* with attribute j yields a square matrix $A_{M \times M}$ for M attributes where r_{ij} denotes the comparative importance of attribute *i* with respect to attribute j. In the matrix, $r_{ij} = 1$ when i = j and $r_{ji} = 1/r_{ij}$.

$$A_{M \times M} = \begin{matrix} \text{Attribute} & 1 & 2 & 3 & --- & --- & M \\ 1 \\ 2 \\ 3 \\ --- \\ --- \\ M \end{matrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & --- & --- & r_{1M} \\ r_{21} & r_{22} & r_{23} & --- & --- & r_{2M} \\ r_{31} & r_{32} & r_{33} & & & r_{3M} \\ --- & --- & --- & --- & --- & --- \\ --- & --- & --- & --- & --- & --- \\ r_{M1} & r_{M2} & r_{M3} & --- & --- & r_{3M} \end{bmatrix}$$

The relative normalized weight (w_j) of each attribute founded by (1) calculating the geometric mean of ith row, this can be represented as,

$$GM_j = \left\{ \prod_{j=1}^{M} r_{ij} \right\}^{1=M} \qquad (3)$$

The row geometric mean method (RGMM) [25] is used to find out the relative normalized weights of the attributes to find out the maximum Eigen value easily and to reduce the inconsistency in judgments; and (2) normalizing the geometric means of rows in the comparison matrix , this represented as follow,

$$w_j = GM_j \Big/ \sum_{j=1}^{M} GM_j \qquad (4)$$

The normalization process performed hereby and determination of significance levels are the third stage. Since the relative importance of different indicators is compared and judged, the fourth stage started because the personal error existed makes the consistency test necessary to make sure if the calculated weight vector value is scientific and rational. The maximum eigenvalue of the judgment matrix calculated as follow,

$$\lambda_{\text{max}} = \frac{1}{M} \sum_{i=1}^{M} \left\{ \sum_{j=1}^{M} r_{ij} w_j \Big/ w_i \right\} \qquad (5)$$

The value of consistency index (CI) obtained from below formula, the smaller of CI is the deviation from the consistency,

$$CI = \frac{(\lambda_{\text{max}} - M)}{(M - 1)} \qquad (6)$$

The random index (RI) for the number of attributes used in decision making. Table IV presents the RI values for different number of attributes.

TABLE IV.   RANDOM INDEX (RI) VALUES

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0	0	0.52	0.89	1.11	1.25	1.35	1.4	1.45	1.49	1.51	1.54	1.56	1.57	1.59

The consistency ratio (CR) calculated to consider the acceptance of the study and it reflects an informed judgment that could be attributed to the knowledge of this case when the value of CR less than 0.1 or equal. Usually, the consistency ratio the ratio of CI and RI, represented as follow,

$$CR = CI/RI \qquad (7)$$

In 2006, Alonso and Lamata have computed a regression of the random indices and propose the formulation [27]:

$$CR = \frac{\lambda_{\text{max}} - M}{(2.7699\,M - 4.3513) - M} \qquad (8)$$

In this case study, Alonson, Lamata used to fit resulting of consistency ratio. The Geometric Consistency Index (GCI) also, used to measure consistency, GCI have been developed in order to discover contradictory judgments, the value of GCI = 0.3147 for M = 3, GCI = 0.3526 for M = 4 and GCI = 0.370 for M > 4 will be accepted based on Aguarón and Moreno-Jiménez determination [28], and it is calculated using:

$$GCI = \frac{2 \sum_{i<j} \ln r_{ij} - \ln \frac{W_i}{W_j}}{(M-1)(M-2)} \qquad (9)$$

The fifth stage in implementation, the overall or composite performance scores for the alternatives obtained by multiplying the relative normalized weight (w_j) of each attribute with its corresponding normalized weight value for each alternative and making summation over all the attributes for each alternative.

$$P_i = \sum_{j=1}^{M} w_j\,(m_{ij})_{normal} \qquad (10)$$

Where (m_{ij}) normal represents the normalized value of (m_{ij}). Pi is the overall or composite score of the alternative Ai. The alternative with the highest value of Pi is considered as the best alternative. Based on the methodology and hierarchical structure in this section, there are 8 pair-wise comparison matrices evaluated by 40 professors from 19 renowned universities in Egypt, in addition, they are responsible about Education System development in ICTP project as MIS managers in their universities. The group comparative matrix for the main criteria, as show in Table V, its analysis illustrated as following,

TABLE V.   THE GROUP COMPARATIVE MATRIX FOR THE MAIN CRITERIA

$$\begin{pmatrix} 1 & 1\,4/5 & 2/5 & 2/5 & 1 & 2\,1/8 & 2/3 \\ 5/9 & 1 & 1/4 & 1/3 & 3/4 & 1\,4/9 & 8/9 \\ 2\,3/5 & 3\,2/3 & 1 & 1 & 2 & 1\,4/5 & 1\,3/7 \\ 2\,1/2 & 3\,1/5 & 1\,1/9 & 1 & 1\,1/2 & 2\,1/2 & 1\,4/9 \\ 1 & 1\,3/8 & 1/2 & 2/3 & 1 & 1\,3/8 & 1\,1/5 \\ 1/2 & 2/3 & 5/9 & 2/5 & 3/4 & 1 & 7/8 \\ 1\,1/2 & 1\,1/8 & 5/7 & 2/3 & 5/6 & 1\,1/7 & 1 \end{pmatrix}$$

The relative normalized weight obtained using Eq. (3) and Eq. (4), w1= 0.118, w2=0.085, w3=0.229, w4=0.228, w5=0.125, w6=0.087, w7=0.0128, the maximum Eigen value λ_{max} computed from Eq.(5), Lambda(max) = 7.186, and consistency index CI = 0.031 from Eq.(6) as shown in same figure , the consistency ratio where random index (RI) of 7 attributes= 1.35 , (CR) value is 0.023 , and GCI is 0.09, by these consistency values , we claims that the feedbacks of the education expertise and decision makers are consistent. After the normalization process in the main-criteria which are seven performance indicators' significance levels are calculated according to the higher education experts. The most important indicators for them in higher education sector have been

identified as teaching & learning indicators, according to the findings teaching & learning indicators have (22.9%), research indicators have (22.8%), internationalization indicators have (12.8%), students support indicators have (12.5%), student's indicators have (11.8%), reputation indicators have (8.7%), and faculty outcomes have (8.5%) importance as shown in Fig.10.

Fig. 10. Analysis and results of main-criteria (indictor's area)

The aggregated group comparative matrix of others levels and iterations calculated and tested in the same way.

V. EXPERIMENTAL RESULTS

After comparison of main-criteria, comparisons and normalizations of performance indicators that are sub-criteria have been carried out. The second, the third and the fourth stages of AHP are repeated herein respectively. Teaching & learning indicators based on sub-criteria in Fig.11, and significance levels calculated after normalization are shown.

Fig. 11. Analysis and results of teaching & learning indicators

Fig.11 determines the significance levels of 7 teaching & learning indicators among themselves and present aggregated group judgment matrix of these sub-criteria. The significance levels of teaching & learning indicators are; faculty to student ratio (26%) ; continuous improvement (17.1%); satisfaction with teaching (survey) (16.1%); graduate employment rate (14.1%); further study (11.4%); faculty to administrative staff ratio (7.6%) and gender balance (7%) .The value of Lambda is 7.177, CR related to this matrix is 0.022 and GCI is 0.08.

Fig.12 determines the significance levels of 4 research indicators among themselves and present aggregated group judgment matrix of these sub-criteria. The significance levels of research indicators are; research excellence and outputs (35.9%); research funding (27.5%); citations per paper (19.3%) and papers per faculty (17.1%). The value of Lambda is 4.025, CR related to this matrix is 0.9 and GCI is 0.03.

Fig. 12. Analysis and results of research indicators

Fig.13 determines the significance levels of 6 internationalization indicators among themselves and present aggregated group judgment matrix of these sub-criteria. The significance levels of internationalization indicators are; recognized accreditations (24.0%); patents (23.5%); international faculty (19.5%); outbound exchange students (13.0%); visiting scholars (10.5%) and international students (9.5%). The value of Lambda is 6.055, CR related to this matrix is 0.09 and GCI is 0.03.

Fig. 13. Analysis and results of Internationalization indicators

Fig.14 determines the significance levels of 3 Students Support indicators among themselves and present aggregated group judgment matrix of these sub-criteria. The significance levels of students support indicators are; low-income outreach (48.0%); environmental impact (survey) (33.3%); and Internal migration (18.7%). The value of Lambda is 3.046, CR related to this matrix is 0.048 and GCI is 0.14.
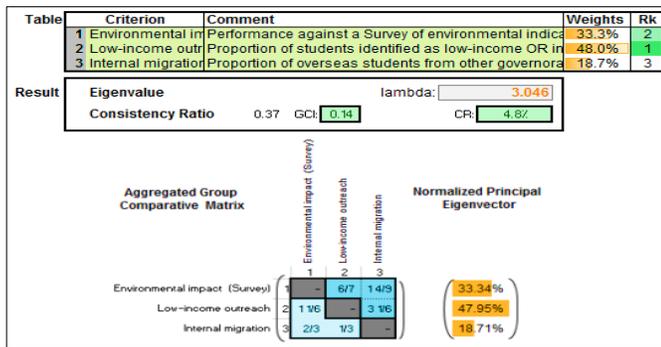
Fig. 14. Analysis and results of students support indicators

Fig.15 determines the significance levels of 6 student outcomes indicators among themselves and present aggregated group judgment matrix of these sub-criteria. The significance levels of student outcomes indicators are; participation (23.7%); overall student satisfaction (survey) (18.7%); completion (16.9%); granted degrees (16.4%); retention rate (15.9%); and scholarships and bursaries (8.4%).The value of Lambda is 6.173, CR related to this matrix is 0.028 and GCI is 0.10.
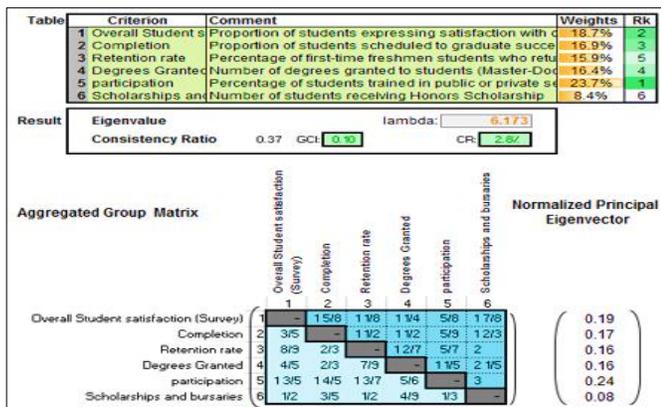
Fig. 15. Analysis and results of Student Outcomes indicators
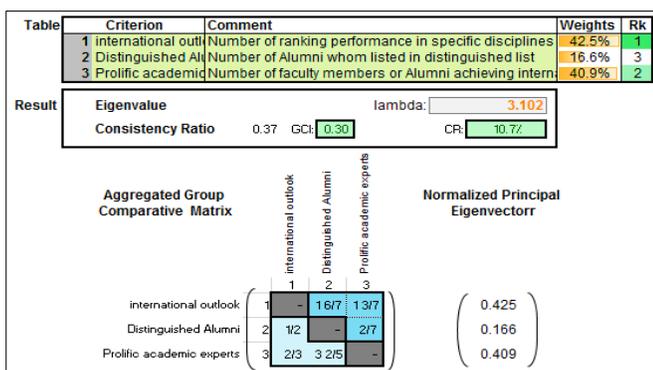
Fig. 16. Analysis and results of reputation indicators

Fig.16 determines the significance levels of 4 reputation indicators among themselves and present aggregated group judgment matrix of these sub-criteria. The significance levels of reputation indicators are; prolific academic experts (40.9%); international outlook (45.0%); and distinguished alumni

(16.6%). The value of Lambda is 3.102, CR related to this matrix is 0.1 and GCI is 0.30.

Fig.17 determines the significance levels of 4 faculty outcomes indicators among themselves and present aggregated group judgment matrix of these sub-criteria. The significance levels of faculty outcomes indicators are; efficiency (30.7%); publications (24.2%); faculty qualifications (23.0%); and attainment (22.1%). The value of Lambda is 4.289, CR related to this matrix is 0.1 and GCI is 0.37.
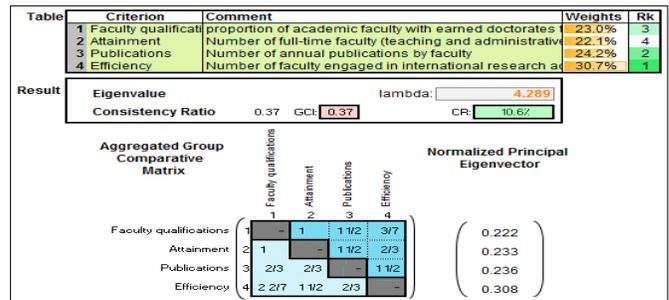
Fig. 17. Analysis and results of faculty outcomes indicators

According to obtained results; in total 33 performance indicators, 7 of which are teaching & learning indicators, 4 research indicators, 6 internationalization indicators, 3 students support indicators, 6 students outcomes indicators, 3 reputation indicators and 4 faculty outcomes indicators, a set of common performance indicators has been built. In the set, the significance level that is a basis to measuring the performance of higher education in Egypt has been designated as well. As a result of product for significance levels related to main and sub-criteria, priorities that are the last stage of AHP have been ascertained to obtain the overall or composite performance using Eq. (10). Table VI is formed in order to understand and apply the set of performance indicators that can be used by the proposed composite index. This table also presents the significance level of each performance indicator per over all sub-criteria. In the indicator Area and weight of area columns in Table VI are 7 indicators areas for higher education systems, KPI's and weights of which are measured and presented in Fig.10. Indicator column represents the performance indicators present in each indicator area on the basis of codes defined in Table III. Weight of type column includes the significance levels all sub-criteria of each KPI. Data Source column represents on which solution or application system can be defined as data source for this kind of KPI or external data source if required. Direction of expectation (DoE) column represents on which direction an expectation, up or down, should occur about the related indicator. Unit of measurement column, explains the unit of value to be obtained by the calculation on the related indicator. Value of weight column is about determining the priorities present in the last stage of AHP. Value weight is obtained by the multiplication of the main criteria and sub-criteria weights. For example, value of weight is obtained by main criteria weight (weight in area of students' outcomes area) * sub-criteria weight (weight in type of retention rate indictor) formulization. 0.118 * 0.159 = 0.0188 = 1.88 %. This calculation is made for 33 performance indicators.

To build the full solution that helps decision makers of higher education in Egypt and domain experts in each university, we have built full integrated solution based-on business intelligence (BI) using Data Warehouse (DW) concepts that will collect required data from Universities Management Information Systems (MIS) as step to extract and calculate required information for the proposed framework.

There are two different database schemes for MIS solutions hosted in universities of Egypt, first scheme related to universities that teaching by Traditional Education System (TES) or Balk Courses Registration (BCR), and other scheme, for institutes that teaching by Credit-Hours Education System (CHES). For that staging database was developed to obtained raw data. Once the data is in the staging area, we encounter staging metadata to guide the transformation and loading processes, including staging file and target table layouts, transformation and cleansing rules, conformed dimension and fact definitions, aggregation definitions, and ETL transmission schedules and run-log results. Moreover, a prototype has been implemented to apply the proposed model and apply comparative analysis using proposed framework, universities in the prototype, anonymously identified as university A, B, C and D as show in Table. VI

TABLE VI.   UNVERSTIEIS THAT INVLOVED IN PROTOTYPE

	Scale	Enrolled Students	Graduate Students	Admitted Students	GRS*
University A	Large	~140 k	~38 k	~34 k	Yes
University B	Large	~130 k	~33 k	~33 k	No
University C	Medium	~50 k	~15 k	~17 k	No
University D	Small	~19 k	~5 k	~4 k	Yes

* State of university that appears in global ranking system

Aggregation method entails the calculation of the ranking of each university according to each individual indicator, e.g. to apply required aggregation method for Granted Degrees indictor under student's outcomes indecator area, we are used Multi-Year Growth Rate Method (MYGRM) to calaculacte this indcator for five acadamic year, by this we can achive Time phased element of SMART cirtira to devlope indictors and to calculate the value of each university for five years, the below steps followed as show in Fig.18 and Fig.21.

1. The equation was known. The equation for annual growth rate percentage over multiple years is
   P =[(f/s)^(1/y)] – 1
   Where f = Final value or population, s = Starting value or population, and y = Number of years
2. The final value was taken and divide it by the starting value.
3. The quotient was raised to the power of 1/y.
4. 1 was subtracted from the product
5. The decimal was converted into a percentage, if necessary, and check your work.

Fig. 18. Steps followed to calculate Aggregation of Granted Degree indictor

Since the all indictor's data aggregated the role of data normalization raised, it is required prior to any data aggregation as the indicators in a data set often have different measurement units, and the normalization methods used in our model show in Table. VII

TABLE VII.   NORMALIZATION METHODS USED IN PROPOSED FRAMWORK

Method	Equation
1. Ranking	$I_{qc}^t = Rank(x_{qc}^t)$
2. Standardisation (or z-scores)	$I_{qc}^t = \dfrac{x_{qc}^t - x_{qc=\bar{c}}^t}{\sigma_{qc=\bar{c}}^t}$
3. Min-Max	$I_{qc}^t = \dfrac{x_{qc}^t - \min_c(x_q^{t_0})}{\max_c(x_q^{t_0}) - \min_c(x_q^{t_0})}$.
4. Percentage of annual differences over consecutive years	$I_{qc}^t = \dfrac{x_{qc}^t - x_{qc}^{t-1}}{x_{qc}^t}$

The normalized and summarized data are maintained in the marts and are organized in a star schema to provide a dimensional view of the data, a dynamic and interactive dashboard with four views were implemented to simulate the composite index functions and represent visual interface for outcomes of campsite index as following, (1) Measures View Dashboard; (2) Indicators View Dashboard; (3) Index View Dashboard; and (4) League View Dashboard, as show in in Fig.19

Fig. 19. Visual Dashboard of Composite Index

As show in Fig.20, the university C ranked as top university, university A ranked as third university. Meanwhile, the university A appears in several Global Ranking Systems. By this we can prove that university C should appears also into Global Ranking Systems but there are some reasons prevented the appearance, and we can conclude these issues as following; (1) related to cut-off, stop displaying university scores after certain rank; (2) related to target audience; (3) related to design and objective; (4) Not aim to label or focus-on Egyptian universities; (5) May lead to redesigning of strategy to improve in the rankings rather than to do what's right for the local setting; and (5)sources of information not exhaustive.

League Table and Z-index for Students Outcome KPIs								
University	Student Satisfaction	Completion	Retention Rate	Degrees Granted	Participation	Scholarships	KPI Proportion (Area)	Maen	Rank
University C	92.544	96.373792033	77.220458425	35.29361624	15.50401352	42.61670631	5.926103445	1	1
University B	76.84	83.085089271	82.036940973	8.026234643	3.651385075	31.95079108	4.260426912	0.173709	2
University A	84.646	67.888991664	81.595618186	7.586081794	6.659542307	9.626227935	3.993649798	0.041369	3
University D	89.936	71.433290847	87.763637565	7.020376162	2.858193409	5.843219305	3.910255348	0	4

Area proportion = (KPI 1 value + ...+ n) * weighted value of area ,   Max- Min , Rank

Fig. 20. Composite Index League Table

VI. Conclusion

In this paper, we proposed a unified and unique composite index for higher education institutions in Egypt aims to, (1) benchmark the overall performance and support self-assessment; (2) transform organizational processes into strategic tools, helping higher education institutions to compare systematically their practice and performance with peer institutions; (3) periodical performance mentoring tool; (4) obtain data to support decision-making; (5) measure performance against and compare with other institutions and assess reasons for any deviations; (6) focus on how the improvement or retracement done rather than who is best; (7) encourage discovery of new ideas through a strategic look inside or outside the institution; (8) follow and set new standards for the institution and higher education system; (9) respond to international benchmarks and performance indicators; (10) set effective targets to achieve accreditation, quality assurance improvement and evaluation processes; and (11) enhance reputation and build new brand for Egyptian case.

Indeed, decision maker and institutes' leaders are forced to make their decisions transparent and comprehensible and draw attention to major issues. The proposed composite index established dependent on the framework [20, 21] than on methodological choices as ranking system, this will make idea of this research usable and can be published.

The AHP algorithm and dynamic dashboard assists our objective judgment and makes the research results more comprehensive and reasonable depending on participation of universities professors as education experts from 19 renowned universities in Egypt. The weights of KPI areas of the proposed composite index belong to the Egyptian situation are summarized into, teaching & learning indicators have (22.9%), research indicators have (22.8%), internationalization indicators have (12.8%), students support indicators have (12.5%), student's indicators have (11.8%), reputation indicators have (8.7%), and faculty outcomes have (8.5%) importance, as green-to-red scale that shown in table VI. And for sub-criteria, the most important indicators have been identified as research excellence & outputs (8.19%) under research indicators area, then Faculty to student ratio (6.1%) under facilities indicators area. No index can be better without; (1) data feeds from different data sources; and (2) using of trends and outcomes as results of this index. For that in the future work, we aims to collect required data from different data sources of higher education institutions in Egypt, for example, MIS and LMS application on each university and linking with other external data sources .

Moreover, enhancement phase of composite index life-cycle will be implemented as feature work and an interactive

dashboard will designed to visualize outcomes of composite index, this will assists our work and place powerful way to track universities' performance and reaching their overall goal.

References

[1] (2013) Human Development Index - Annual Report 2013 [Online]. Available: http://hdr.undp.org/en/countries/profiles/EGY

[2] (2013) Supreme Council of Universities – Egypt strategy plan for education until 2021 [Online]. Available: http://mhe-spu.org/site-old/admin/uploads/events/files/Report_on_Master_Plan.pdf

[3] M. El-Hadidi, H. Anis, S. El-Akabawi, A. Fahmy, M. Salem, A. Tantawy, A. El-Rafie, M. Saleh, T. El-Ahmady, I. Abdel-Moniem, A. Hassan, A. Saad, H. Fahim, T. Gharib, M. Sharawy and K. Abdel-Fattah, "Quantifying the ICT Needs of Academic Institutes Using the Service Category-Stakeholder Matrix Approach," *6th ICICT'08 Conference Proc*, December 2008, pp. 107-113.

[4] (2012) Information and Communications Technology Project , unified Management Information System brief description [Online]. Available: http://www.ictp.org.eg/index.php/ar/2012-11-03-22-55-31/2012-11-03-22-56-46

[5] Liu, Nian Cai, Qi Wang, and Ying Cheng, eds. Paths to a world-class university: Lessons from practices and experiences. Vol. 23. Springer, 2011.

[6] Alexander, F. King. "The changing face of accountability: Monitoring and assessing institutional performance in higher education." Journal of Higher Education (2000): 411-43.

[7] McKinnon, K. R., S. H. Walker, and D. Davis. "A manual for Australian Universities." Australian Higher Education Division, Australia (2000).

[8] Sadlak, Jan, and Nian Cai Liu. The world-class university and ranking: Aiming beyond status. Bucharest, Romania/Shanghai, China/Cluj-Napoca, Romania: Unesco-Cepes, 2007.

[9] Hazelkorn, Ellen. "How Rankings are Reshaping Higher Education." (2013).

[10] Rauhvargers, Andrejs. "Global university rankings and their impact" Leadership for WorldClass Universities Challenges for Developing Countries June (2011).

[11] Zhou, P., B. W. Ang, and D. Q. Zhou. "Weighting and aggregation in composite indicator construction: a multiplicative optimization approach." Social Indicators Research 96, no. 1 (2010): 169-181.

[12] Ismail, Engr Muhammad. "Ranking of universities." In 2nd International Conference on Assessing Quality in Higher Education, pp. 6-8. 2010.

[13] Saaty, T. L. (1988). What is the analytic hierarchy process? (pp. 109-121). Springer Berlin Heidelberg.

[14] Lingxi, Zhu, and Yu Hao. "Study on the AHP-based evaluation system of excellence of university classes." In Education and Management Technology (ICEMT), 2010 International Conference on, pp. 242-247. IEEE, 2010.

[15] Badri, Masood A., and Mohammed H. Abdulla. "Awards of excellence in institutions of higher education: an AHP approach." International Journal of Educational Management 18, no. 4 (2004): 224-242.

[16] Lavy, Victor. "Evaluating the effect of teachers' group performance incentives on pupil achievement." Journal of political Economy 110, no. 6 (2002): 1286-1317.

[17] Saltelli, Andrea. "Composite indicators between analysis and advocacy." Social Indicators Research 81, no. 1 (2007): 65-77.

[18] Bandura, Romina. "A survey of composite indices measuring country performance: 2008 update." Office of Development Studies, New York: United Nations Development Program (2008).

[19] Giovannini, Enrico, Michela Nardo, Michaela Saisana, Andrea Saltelli, A. Tarantola, and Anders Hoffman. Handbook on constructing

composite indicators: methodology and user guide. Organisation for Economic Cooperation and Development (OECD), 2008.

[20] Saisana, Michaela, Béatrice d'Hombres, and Andrea Saltelli. "Rickety numbers: Volatility of university rankings and policy implications." Research Policy 40, no. 1 (2011): 165-177.

[21] Paruolo, Paolo, Michaela Saisana, and Andrea Saltelli. "Ratings and rankings: voodoo or science?." Journal of the Royal Statistical Society: Series A (Statistics in Society) (2012).

[22] Saaty, Thomas L., and Luis G. VARGAS. "Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process. The Analytic Hierarchy Process, vol. 6." (2000).

[23] Eccles, Charles. "The use of university rankings in the United Kingdom." Higher Education in Europe 27, no. 4 (2002): 423-432.

[24] XU, Qian, Wen-lei QU, and Fang-chun QIU. "The design for the index system of college teachers performance assessment based on the AHP method [J]." Journal of Qingdao University of Science and Technology (Social Sciences) 4 (2008): 021.

[25] Forman, Ernest, and Kirti Peniwati. "Aggregating individual judgments and priorities with the analytic hierarchy process." European Journal of Operational Research 108, no. 1 (1998): 165-169.

[26] Escobar, María Teresa, and José María Moreno-jiménez. "Aggregation of individual preference structures in AHP-group decision making." Group Decision and Negotiation 16, no. 4 (2007): 287-301.

[27] Alonso, Jose Antonio, and M. Teresa Lamata. "Consistency in the analytic hierarchy process: a new approach." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 14, no. 04 (2006): 445-459.

[28] Aguaron, Juan, and José María Moreno-Jiménez. "The geometric consistency index: Approximated thresholds." European Journal of Operational Research 147, no. 1 (2003): 137-145.

[29] Mohamed Rashad M Elhefnawy, Ali El-Bastawissy, Mona Kadry (August, 2014). "Towards a Composite Index for Measuring the Higher Education Institutions in Egypt. "Science and Information Conference (SAI), London, 2014, pp. 31-41. IEEE, 2014.

Fig. 21. Aggregation simple for Granted Degree indictor

TABLE VIII. FULL COMPOSITE PERFORMANCE SCORES FOR COMPOSITE INDEX FRAMEWORK

#	Area	Weight of Area (%)	Code	Weight In Type (%)		Indicator	Data Source	DoE	Unit of Measurement	Value of Weight (%)
1	Students Outcome	11.80%	S1	18.69%	▲	Overall Student satisfaction (Survey)	Survey DB	↑	Per Cent	2.21%
2			S2	16.89%		Completion	MIS DB	↑	Per Cent	1.99%
3			S3	15.95%	=	Retention rate	MIS DB	↓	Per Cent	1.88%
4			S4	16.38%	=	Degrees Granted	MIS DB	↑	Count	1.93%
5			S5	23.73%	▲	participation	MIS DB	↑	Per Cent	2.80%
6			S6	8.36%	▼	Scholarships and bursaries	MIS DB	↑	Count	0.99%
7	Faculty Outcome	8.54%	F1	22.20%	▼	Faculty qualifications	MIS DB	↑	Per Cent	1.90%
8			F2	23.34%	▼	Attainment	MIS DB	↑	Count	1.99%
9			F3	23.62%	▼	Publications	MIS DB	↑	Count	2.02%
10			F4	30.84%	▲	Efficiency	MIS DB	↑	Count	2.63%
11	Teaching & Learning	22.91%	T1	26.64%	▲	Faculty to student ratio	MIS DB	↓	Fraction	6.10%
12			T2	7.58%	▼	Faculty to Adminstrative staff ratio	MIS DB	→	Fraction	1.74%
13			T3	16.12%	=	Satisfaction with teaching (Survey)	Survey DB	↑	Per Cent	3.69%
14			T4	11.42%	▼	Further study	MIS DB	↑	Per Cent	2.62%
15			T5	7.04%	▼	Gender balance	MIS DB	→	Fraction	1.61%
16			T6	14.12%	=	Graduate employment rate	MIS DB	↑	Per Cent	3.24%
17			T7	17.08%	=	Continuous improvement	MIS DB	↑	Count	3.91%
18	Reserch	22.79%	R1	19.38%	▼	Citations per paper	Extrnal DS	↑	Per Cent	4.42%
19			R2	17.12%	▼	Papers per faculty	MIS DB	↑	Per Cent	3.90%
20			R3	35.93%	▲	Reserch excellence & outputs	Extrnal DS	↑	Count	8.19%
21			R4	27.57%	=	Research funding	ERP DB	↑	Per Cent	6.28%
22	Students Support	12.49%	A1	33.34%	=	Environmental impact (Survey)	Survey DB	↑	Per Cent	4.16%
23			A2	47.95%	▲	Low-income outreach	MIS DB	↑	Per Cent	5.99%
24			A3	18.71%	▼	Internal migration	MIS DB	↓	Per Cent	2.34%
25	Reputation	8.70%	E1	42.55%	▲	international outlook	Extrnal DS	↑	Count	3.70%
26			E2	16.58%	▼	Distinguished Alumni	Extrnal DS	↑	Count	1.44%
27			E3	40.88%	▲	Prolific academic experts	Extrnal DS	↑	Count	3.55%
28	Internationalisation	12.77%	I1	19.47%	▲	International faculty	MIS DB	↑	Per Cent	2.49%
29			I2	10.46%	▼	Visiting scholars	MIS DB	↑	Count	1.34%
30			I3	9.54%	▼	International students	MIS DB	↑	Per Cent	1.22%
31			I4	13.00%	▼	Outbound exchange students	MIS DB	↑	Per Cent	1.66%
32			I5	24.00%	▲	Recognized accreditations	MIS DB	↑	Count	3.06%
33			I6	23.53%	▲	Patents	MIS DB	↑	Count	3.00%
	Total (%)=	100.00%							Total (%)=	100.00%

New Approach of Using Structural Modelling for Personalized Study Planning

Raita Rollande*

Engineering Research Institute
"Ventspils International Radio
Astronomy Centre" of Ventspils
University College,
Ventspils, Latvia

Janis Grundspenkis

Department of Systems Theory and
Design,
Riga Technical University*,
Riga, Latvia

Antons Mislevics

Department of Systems Theory and
Design,
Riga Technical University,
Riga, Latvia

Abstract—**The planning of individual studies becomes more and more topical. Alongside the fast and diverse rhythm of life, people need individual approach to study planning. It ensures wider availability of learning for different social groups. The use of information and communication technologies (ICT) in the process of studies makes it more attractive and more interesting as well as more adequate to the demands of the 21st century. The authors have developed graph based framework for personalization of education process based on the set of four graphs representing the structure of the study programme, the structure of the study courses, the concept maps, and the learning objects [1, 2, 3, 4, 5]. Special tool is needed which allows the learners themselves to design the learning strategy corresponding to their interests. The authors had already implemented personalized study planning process in a prototype, which allows to create a personalized study programme, and then to plan the course learning, setting the courses in the required sequence [3, 4, 5]. In this paper the authors propose the use of the methods of structure analysis to calculate the ranks for the node of the graphs thus detecting the most significant nodes in the graph structure. The calculations for the ranks are made for the graph of the study programme, for the graph of study courses, and for the concept maps. The calculation of ranks for the graph nodes allows detecting the most significant courses in the study programme, the most important topics in the study course and the most essential concepts in the concept map.**

Keywords—intelligent tutoring system; personalized education; graphs; structural modelling; element ranking and structural analysis

I. INTRODUCTION

According to the analysis of the planning documents of the field of education [7, 8, 9, 10, 11] there is need for the educational system that offers individual approach to the learner. The researched documents of planning showed that the learner should be ensured by larger variety of study forms and more personalized studying to satisfy the need for new competences and skills. The full-time forms of training do not always correspond to the learner's demands because he/she has individual demands and interests for studies as well as different skills, background knowledge, learning style, and besides the learner very often studies same time working. The use of ICT in the process of studies has significant role as it offers electronic study materials. ICT can be used as additional tools in the organization of classes, in creation of study materials, in knowledge assessment, in communication, and so on. ICT

allow to make the studying process more attractive and interesting as well as more adequate to the demands of the 21st century [12]. The use of ICT in studies and for storage the electronic materials is ensured by computerized studying systems which offer to the learner to study in the environment that does not depend on time and place. However, it should be mentioned that computerized studying systems do not ensure individual approach. Thus it is necessary to have such studying systems which implement personalized learning and it starts with creation of individual study plan according to the learner's wishes and needs. Personalized learning can be carried out in two ways: the lecturer cooperates with the learner individually or there is developed intellectual learning system. The authors had already implemented graph-based framework for personalization of education process. There are used graphs to represent the structure of the personalized study planning framework. Those allow not only showing the structure in a transparent way, but also allow using the methods of structure analysis to calculate the ranks for the nodes of the graphs thus detecting the most significant node in the graph structure. In this paper, the authors perform study programme and course structure graph analysis using such structural modelling method as calculating ranks.

The paper is organized as follows. The next part is the existing personalized learning construction systems. The third section presents Graph based framework for personalization of education process. The fourth section shows personalized study planning system's SPS prototype. The fifth section describes SPS analysis using structure modelling methods. The sixth section describes graph structure analysis. The paper concludes with a short summary and it outlines directions for the future research.

II. PERSONALIZED LEARNING CONSTRUCTION SYSTEMS

This chapter is devoted to the existing personalized learning construction systems.

Kabicher and Motschnig-Pitrik [13] have elaborated the tool for study programme visualization using the CEWebS (Cooperative Environment Web Services) platform, which is designed to visual description of study programme using the graph. CEWebS is intended only for teachers; the learners cannot work with it independently. Besides the study programme showed in the system is static not dynamic, i.e., all have only one tutoring scenario.

Gestwicki [14] and Toombs [15] elaborated the tool for visualisation of the study programme CurricVis. The tool automatically generates the study programme as oriented graph based on information present in the knowledge basis. CurricVis is the design of the programme that presents the structure of the study programme in a form of graph. In comparison with the previous tool, CurricVis allows to be operated not only by the designers of the study programme but also by learners themselves thus enabling them to participate in the study programme planning process. CurricVis tool does not have an interactive graph creation mode.

Zucker in the article [16] offers the tool ViCurriAS that includes two modules. Study programme module is for the design and the modification of the study programme. The module allows the faculty methodologists and consultants arrange the courses in the study programme and define the pre-requisitions of these courses. The Consultation module allows the consultants to enter the marks or planning information in the same graph that was created in the Study programme module. ViCurriAS system same as the two previously described systems does not offer to the learner him/herself to create the study programme dynamically neither realizes further training processes.

A different approach from the previously described tools is offered by Auvinen [17], unlike the previously described systems this system has described relationship not between the study courses but between study outcomes. He offers the design of study programme using the graphs of study outcomes. The construction of the study programme is based on choice of competences. According to the chosen competences the graph of acquirable study courses relating to study results is created. In this system the learner is able to design his/ her own study programme. As negative trend should be regarded the thing that the learner may ignore topics that are related to study outcomes which are not related to his/her competences.

Nkambou, Gauthier and Frasson [18, 19, 21, 21] offer the tool CREAM for designing the study programme. The study programme creation environment allows automatically generate the study programme. The course creation environment allows to create courses with the given parameters. At the basis of CREAM tool study programme is the module of skills. The learner chooses the acquirable skills and after that CREAM tool generates suitable study programme. In CREAM tool as in previous tools study courses are described in the form of graph.

As existing personalized study planning systems do not ensure the learner to create personalized study plans and courses within one system it is necessary to have supplements in the pedagogical module that support personalized learning allowing the learner him-/her-self design the study plan suitable to one's needs, see the topics of the study course and choose the sequence of study course acquisition, get acquainted with concept connection and choose learning objects of topics.

According to the analysis of the study system architecture, there follows such demands for the new architecture:

- develop a united system which allows to personalize the study process beginning with design of study programme plan to its acquisition;

- describe the study programme, study courses, topics, concepts and learning objects;

- include the possibility for the learner to personalize the study process allowing to develop the individual study plan, set the sequence of course acquisition, frame the concept map and choose the learning objects for acquisition of topics and concepts;

- the structure of study programmes, study courses, concepts and learning objects displayed in the graph mode;

- study courses and course topics define the range of prerequisites and restrictions so that the learner himself could develop individual study programme plan and set the sequence of topic acquisition;

- offer the designing of study programme plan for lifelong learning courses;

- ensure the designing of study programme plan taking into account that the learner can participate in mobility programme or study after moving from another study programme, or after the acknowledgment of credit points;

- offer several types for study programme, study course, concepts and learning object graph visualization so that the learner could get transparent structure of study programme or study course with different amount of nodes;

- include the determination of significance of graph nodes in order to help the programme directors and teachers manage the structure of study programme and course as well as to help learners in planning personalized studies.

The next chapter describes the graph based framework for personalization of education process, which consists of four graphs.

III. GRAPH BASED FRAMEWORK FOR PERSONALIZATION OF EDUCATION PROCESS

The previous chapter described different learning systems which allow to develop study programmes and courses to achieve personalized study planning. Rather seldom there are systems which allow the learner to design both the study progamme plan and courses, but there are no systems at all [3], which allow to fulfill the whole learning scenario beginning with the designing of the study programme plan and ending with the choice of learning objects for each concept and topic be learned as well as defining the most important courses, topics and concepts. As the result in the previous chapter there are defined demands for the development of such individual learning system. In this chapter there is offered the framework of personalized study planning which is based on the set of graphs.

To fulfill the personalized study planning from creation of study programme plan to the choice of learning objects, there is developed the framework of personalized study planning based on the following set of graphs (Fig. 1) [1, 2, 3, 4, 5]:

- a graph representing a conceptual structure of study program $G_1(V_1,Q_1)$ allows to design individual study program;

- a graph representing study course $G_2(V_2,Q_2)$ allows to develop individual learning scenario;

- a graph visualizing each topic using concept map $G_3(V_3,Q_3)$ ensures mapping of each topic to the corresponding concept map;

- a graph representing learning objects $G_4(V_4,Q_4)$ describes each concept with learning objects.
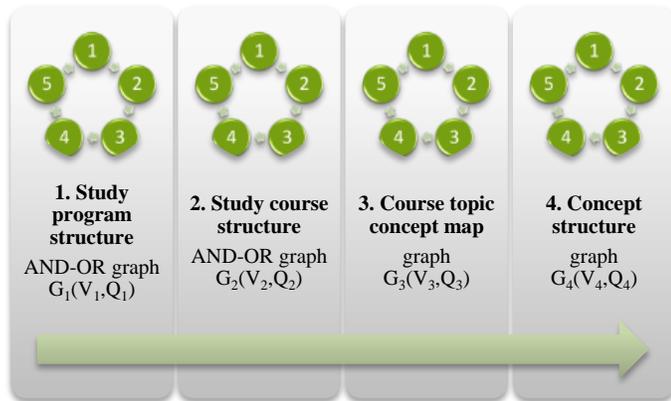
Fig. 1. Graph based framework for personalization of education process [3]

The graphs of personalized study planning framework are related. Each next graph results from the previous. Graph G_1 allows the learner to create personalized study programme choosing the courses to be included into one's individual study programme. After creation of study programme the learner chooses the study courses which he/she wants to master. To describe the course structure is meant graph G_2, which shows the topics that are included in each study course. To master the topics of the study course the next level of detail uses graph G_3, which describes the concepts of each topic and their mutual relationship. Finally the fourth and the concluding graph of the study planning framework allows the learner to personalize mastering of each topic or concept by choosing the learning objects, and that corresponds graph G_4.

To use the personalized study planning framework the following steps should be fulfilled for the development of personalized study planning framework (Fig. 2):

- Step 1 – at first graph G_1 is designed and it starts with definition of study programme which is the node of the graph G_1 root;

- Step 2 – in the next level are defined the study years which are the direct successors of the root node;

- Step 3 – in the next level for each study year adjust two successor nodes – spring semester and autumn semester;

- Step 4 – for each semester in the next level of graph G_1 define successor nodes which correspond to the tittle of the study programme part;

- Step 5 – for each study programme part in the next level as the nodes of successors define respective tittles of the study courses;

- Step 6 – for each course tittle as predecessors describe tittles of other courses which are prerequisites for mastering the course;

- Step 7 – start to design graph G_2 and define the study course tittle and its ECTS volume which is one of the graph G_2 root nodes. The number of G_2 corresponds to the number of course tittles described in graph G_1;

- Step 8 – define for the study course node the successor nodes – numbers of the classes, their number corresponds the volume of credit points, one ECTS is 12 classes, one class is 2 academic hours;

- Step 9 – for each number of the class in the next level of the graph successor nodes define the tittles of the topics which are mastered in definite class;

- Step 10 – for each topic tittle define predecessors which are shown with the predecessors' node and which shows the tittles of the prerequisite topics;

- Step 11 – choose the tittle of the topic and define its concepts which correspond graph G_3, as the result the number of graphs G_3 corresponds the number of topics;

- Step 12 – relate the concepts with links and define the semantics of the links;

- Step 13 – choose the tittle of the topic from graph G_2 or concept from graph G_3, which is the node of the root of graph G_4. The number of graphs G_4 corresponds the number of topics of graph G_2 or the number of concepts of graph G_3;

- Step 14 – for each root node of graph G_4 in the next level as successors' node define the tittles of learning objects;

- Step 15 – for each tittle of the learning object define learning objects.
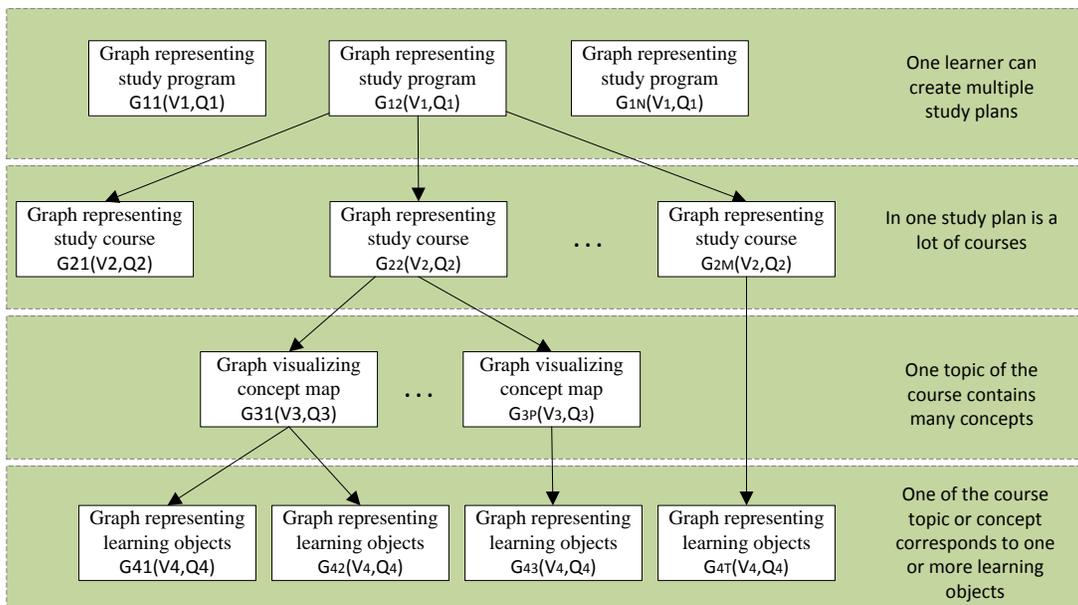
Fig. 2. General structure of personalized study planning framework

This framework allows any student to tailor a study programme by adapting the modularized curriculum structure and to choose the suitable learning strategy for each study course. Next section describes personalized study planning system prototype SPS.

IV. PERSONALIZED STUDY PLANNING SYSTEM SPS PROTOTYPE

The previous section describes the concept of ITS - graph-based framework for personalization of education process. This section describes developed personalized study planning system's SPS prototype.

In SPS prototype is implemented system data basis (*SPSDB*), data view and insert forms (*SPS.DataManager*) and study programme construction software application (*SPS.StudentStudyProgramBuilder*). SPS prototype has been elaborated using *Microsoft .NET Framework 4.5* software creation platform [22, 23]. There are used such built-in technologies as:

- *Windows Presentation Foundation (WPF)* [24] – for creation of graphic user interface;

- *LightSwitch* [25] – for creation of data review and types of input.

SQL Server 2012 Express LocalDB [26] is used as the system of database management for the local database. It simplifies the installation of prototype because there is no need to install and configure the database server. As SPS does not use specific possibilities of *SQL Server 2012 Express LocalDB* the solution can be used also for other versions of *SQL Server* [27].

The prototype was elaborated using the programme creation tool *Visual Studio 2012* [28]. Implementing the operations with graphs and graph visualizations open code software libraries *QuickGraph* [29] and *Graph#* [30] are used.

Fig. 2 shows the components of the prototype describing also the technologies used in their implementation.
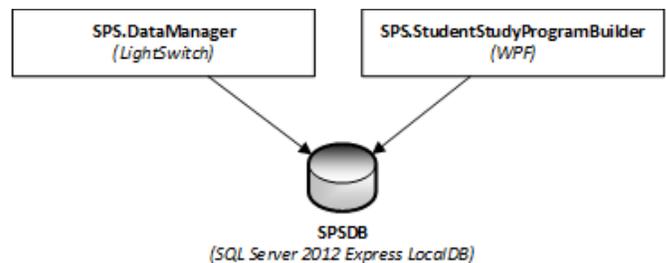
Fig. 3. SPS architecture [4, 5]

Personalized study planning system SPS prototype contains two tools [4, 5]:

- *SPS.DataManager*, which is provided to input data on study course field, study programme, parts of the study programme, teachers, courses, the structure of course topics, course topics and learners, and student study programmes in order to fulfil personalized tutoring (shown in Fig. 4);
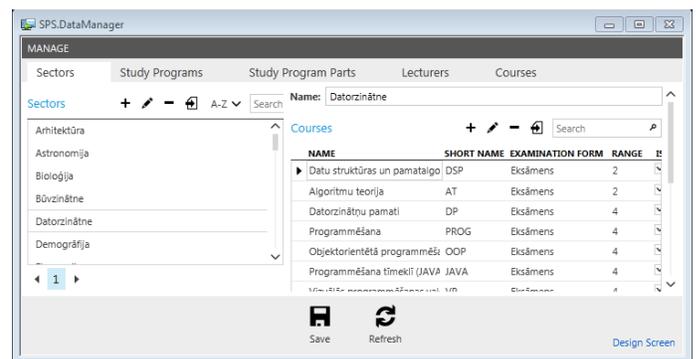
Fig. 4. SPS DataManager [4, 5]

- *SPS.StudentStudyprogramBuilder*, which is provided to determine personalized study programme (see Fig. 5) and the sequence of acquisition of study course topics (see Fig. 6).

Fig. 5. SPS.StudentStudyprogramBuilder personalized study programme
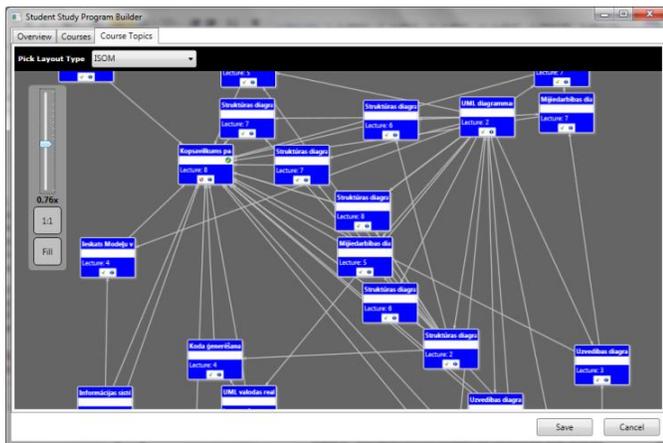
Fig. 6. SPS.StudentStudyprogramBuilder personalized study course acquire [4, 5]

For the personalized study planning system there are distinguished four groups of users: learners, study programme directors, teachers, study programme administrators.

The next chapter describes SPS structure analysis using structure modelling methods.

V. STRUCTURE ANALYZE

The analysis of graph structure is described in several papers [31, 32, 33]. The authors have chosen to use offered in paper [31] approach of structure analysis – structure modelling because it allows to define the significance of elements which is the main difference from other methods. Structure modelling is the way of topological modelling based on computerized construction and analysis of models, development of knowledge basis, and the use conclusion procedure [31]. It started to develop in mid 1970s and an essential investment in its development has been contributed by 13 researchers of Riga Technical University [31]. It is designed technical system with physically multiple elements for mathematical modelling in the circumstances of imperfect information [31]. In structure modelling there are investigated relationships between structure elements, the importance of elements in functioning of the system as well as the assessment of consequences in case of element elimination [31].

The use of structure modelling approach for the analysis of the structure of study programme, study courses, and concept maps is new its application. So far it was more oriented towards the structure analysis of complicated systems with physically multiple elements, and technical diagnostics. Structure modelling has several methods that allow to perform structure analysis, judge about the role of elements in the structure, and the common characteristics of the structure. One of the methods which is included in the structure analysis is the calculation of the ranks of graph nodes. The acquisition of the ranks allows to define the degree of importance of the node in the common structure. The higher is the rank of the element the closer this element is related to other elements in the structure and the more serious consequences may arise if it is excluded from the structure [31].

In structure modelling it is distinguished between qualitative and quantitative structure analysis [31]. Qualitative analysis defines the importance of the graph nodes. Quantitative analysis uses the distance between the elements defined in the graph theory. As it is essential to define the

importance of graph nodes in the elaborated framework graphs G_1, G_2 and G_3, from G_1 obtaining the most important course in the study programme, from G_2 the most important themes in the course, and from G_3 the most important concepts, the authors use the qualitative analysis of the structure. In order to make qualitative analysis of the structure, ranks should be calculated for the nodes [31]. There are chosen three methods for rank calculations:

- By the local level of the node R^1_{LP}. To determine the rank of the elements of the local level, element input and output nodes have to be defined, and after that the sum of input and output nodes is calculated by which the elements are ranked.

- By the number of routes in the graph which contain the given node R^1_{CE}. In this case wider analysis of node mutual relationship is carried out, stating in how many different routes the node is included. In order to do that, first is stated the set of all routes between input and output nodes. After that the number of routes containing the given node is calculated. The obtained number of routs for each node is divided with total number of routes found. The highest rank is given to the nodes with the greatest value which is obtained dividing the number of routes for each node to the total number of routes in the graph. The rank R^1_{CE} shows the structural importance of the node, and it reflects in how many learning scenarios the course, topic or concept is included.

- By the number of attainable nodes R^2_{CE}. In this case there are taken into account all routes that make output from node, but the routes making input into the node are not taken into account. In order to find the rank according to the number of achievable nodes, in the matrix line of attainability all the elements are added and the attainability component is obtained, which afterward is divided to total number of nodes in the graph [31]. The greater is the value, the higher is the rank.

To calculate the element structural significance, firstly the summary ranking R_{sum} is calculated:

$$R_{sum}(i) = R^1_{LP} + R^2_{CE} \text{ [31]} \qquad (1)$$

Secondly, the summary ranks are arranged by places and thus a total rank of elements R_{tot} is obtained. Then the element's structural significance N(i) can be calculated:

$$N(i) = 1 + \frac{1 - R_{tot}}{R_{max}}, \text{ [31]} \qquad (2)$$

where N(i) – element's structural significance;

R_{tot}– element's total rank;

R_{max} – the maximum value of the sum of rank.

To calculate the node structural significance usually there are used two ranks: by the number of routes in the graph and

by the number of attainable nodes [31], and the authors offer for calculation of the node structural significance also to include the third rank which is calculated by the local level of the node. If the nodes are ranked according to their local levels, then there are analyzed direct links, but indirect links that are essential in complicated systems, are ignored [31]. Thus the local level analysis in complicated systems is not actual. However, it is different in personalized study planning system where calculation of ranking by the local levels of nodes is essential when it is necessary to analyze the local information of each study course, topic, or concept separately. To calculate the values of node structural significance which is based on 3 rank values, first is calculated the summary rank R_{sum}^1 taking into account 3 ranks.

$$R_{sum}(i)^1 = R^1_{CE} + R^2_{CE} + R^1_{LP} \qquad (3)$$

After that to calculate the values of nodes structural significance N(i)[1], that is based on 3 rank value, the following formula is used:

$$N(i)^1 = 1 + \frac{1 - R_{tot}}{R_{max}}, \qquad (4)$$

In order to analyze the study programme and the course topic graph structure in personalized study planning system SPS, ranks are calculated for each element of the graph. This is performed in personalized study planning system SPS tool SPS.StudentsStudyProgramBuilder by going to tab Courses with the mouse cursor showing the course then open the information window (Fig. 7), which shows calculated ranks:

- Rank by the local element level R^1_{LP} (in Fig. 7 marked as R1);

- Rank by the number of routes in the graph which contain the given node R^1_{CE} '(in Fig. 7 marked as R3);

- Rank by the attainable number of vertices R^2_{CE} (in Fig. 7 marked as R2);

- Element's total rank R_{tot} (in Fig. 7 marked as Rtot);

- Element's structural significance N(i) (in Fig. 7 marked as N).
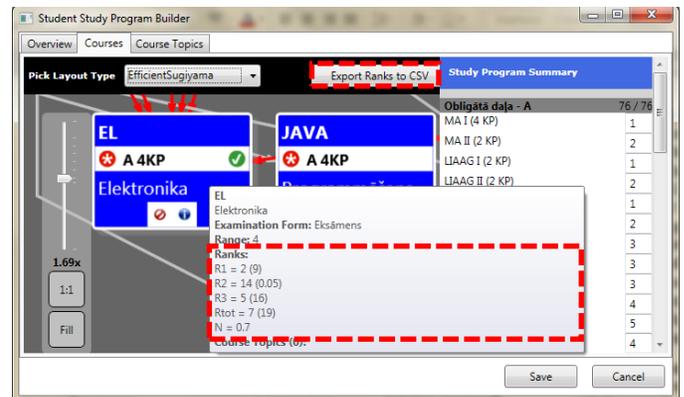
Fig. 7. Personalized study programme course window

To get an overview of calculated rank values for all courses in Microsoft (MS) Excel file, the option <Export Ranks to CSV> should be used (Fig. 7).
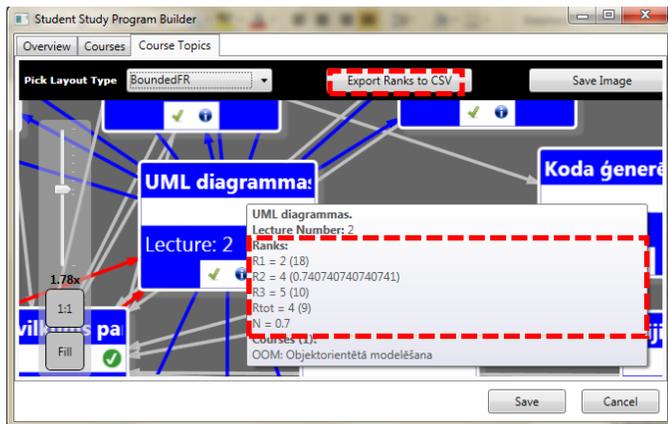
Fig. 8. Personalized study programme study course topic window

In order to analyze the course topics graph structure, similar steps should be performed in Course Topics tab of the tool (Fig. 8). As a result course topic's structural analysis is calculated by the same methods as the structural analysis of study programme graph.

After examining a summary table of structural analysis for the study programme graph of personalized study planning system, it can be concluded:

- By structure analysis of the local levels, it can be seen, that the highest rank is assigned to the course 'Programming', further followed by courses 'Mathematical analysis I' and next courses with equal rank are 'Information system analysis and design' and 'Electronic'. After that other courses follow. Thus are obtained the most significant courses of the study programme by their direct links with other courses. It means also great concentration of learners mastering these courses hence it is advisable to divide such courses into smaller groups so that such courses are not overloaded with too big number of learners. The exclusion of courses with the highest rank from the graph threatens the realization of the study programme because such study courses are prerequisites for other courses. If the learner does not master or masters badly courses with high ranking value, he/she will not possess the necessary knowledge for mastering the related courses.

- Calculating the ranks by the number of routes in the graph which contains the given node, it could be concluded that the higher is the rank, in more learning routes the course is included. Study courses which are included most in different learning routes are 'Basis of Computer Sciences', 'Mathematics Analysis I', 'IS Analysis and Design', and 'Intellectual Network and Computer Telephony'. With equal ranking values follow 'Electronics' and 'Data Processing Systems'. Using the given rank calculation type, higher ranks are obtained by output of the system or their closest

predecessors, this statement is valid with courses 'Electronics' and 'Data Processing Systems'. If the learner does not master or masters badly these courses, he/she will not possess the necessary knowledge for mastering the whole study route. If the programme director want to exclude any high ranking course then it is advisable to analyze the whole study route before making decision.

- When analysing the programme graph structure by the number of attainable nodes, the most important courses are 'Basis of Computer Sciences', 'Programming' and with the same rank value follow courses 'Mathematical Analysis I' and 'Mathematical Analysis II'. After that other courses follow. This rank allows to identify the learning routes for mastering the study plan which is out-going from the study course, thus determining the course branching. The higher is the rank for the course the more essential is the role of it in the mastering other courses that are included in the route of this course. If the learner masters insufficiently or does not master at all the courses which have high ranks by the number of attainable nodes, then he/she cannot reach the study programme learning outcomes. High rank by the number of attainable nodes allows to conclude that the learner has to master the course carefully because that influences mastering the other courses. It is advisable to divide the learners into smaller groups in full-time studies to help the learner to master such course and so that the lecturer could work in closer contact with learners. It is advisable to attract substitute lecturers who can substitute the main lecturers of the course if needed and thus ensuring continuous knowledge transfer to the learners. In training systems it is advisable to revise the content of those study courses which have high ranks by the number of attainable nodes and offer additional examples to ensure mastering the course on required quality.

- The total rank sums up R^1_{CE} and R^2_{CE}, in resulting assessment a total importance of the node. Assessment of the results of the elements total rank, the most important courses are 'Programming', 'Mathematical Analysis I' and courses of the same rank value followed 'Mathematical Analysis II' and 'Basis of Computer Sciences'. It is advisable that study program's directors determine those courses as compulsory mastered courses and the learners have to obligatory include them in their study programme. Consequently, if the study program's director wants to replace these courses with other courses, he/she needs to analyse the next level course topic correlation to the new course to integrate the most important current topics of the course.

The structural significance value of nodes $N(i)^1$ is calculated using the ranks by local levels, by the number of routes in the graph and by the number of attainable nodes. When compare the calculated structural significance value of nodes by two ranks ($N(i)$), with the calculated structural significance value of nodes by three ranks ($N(i)^1$), there is observed the difference in the results. According to $N(i)^1$ the 1st place is shared by the course 'Basis of Computer Sciences',

which according to N(i) also is in the 1st place, and course 'Mathematics Analysis I', which according to N(i) is in the 2nd place. The significance value of the course 'Mathematics Analysis I' has increased because this course had the second highest rank by local level of nodes. According to $N(i)^1$ in the 2nd place is study course 'Programming' the value of which is increased by two places in comparison with N(i) calculated value, because this course had the first highest rank by local level of nodes. The 3rd place according to $N(i)^1$ is shared by the course 'Linear Algebra and Analytical Geometry I', which by N(i) also was in 3rd place, and the course 'Mathematics Analysis II', which by N(i) was in the 4th place. It should be noted that N(i) and $N(i)^1$ values are similar, and changes of $N(i)^1$ values in correlation to N(i) influence the calculated rank values by local level. The author of the paper considers that to evaluate the graph structure is more useful by the structural significance value of nodes which are obtained taking into account all three ranking values because, for example, the course 'Electronics' share the 2nd place in rank calculations by local level although in ranking calculation by the number of the routes it ranks the 5th, but in ranking calculation by the number of attainable nodes it is only in 14th place. In order to determine the structural significance of this course it is important to know not only in how many routes this course has been included or how many different routes are possible from this course but it should be taken into account with how many courses this course is directly linked. Every calculation of node rank has its own meaning, but to obtain the structural significance value of node it is necessary to take into account all 3 ranks. Accordingly structural significance values of nodes, based on 3 ranks are calculated also for the study course structure graph G_2 and for the concept map G_3.

The calculation of the structural significance value in study programme graph G_1 helps to divide the study courses according to the parts of the programme, i.e. study courses with higher structural significance values can be added to part A, with low structural significance values to part C, and the rest of the remaining courses that are in the middle add to part B. In order to determine the limits of how many courses to include in each part of the programme, the volume of ECTS can be taken into account.

Examining the structure analysis summary (Annex 1) personalized learning planning system programme course topic graph shows that:

- By structure analysis of the local levels, it can be seen that the highest rank is assigned to the topic 'Overview of the course object-oriented modelling', 'UML diagrams', 'Structure diagram: The Class diagram' and with equal rank values follow 'Unified Modelling Language - UML' and 'System dynamic model'. After that other course topics follow. Exclusion of the topic of the highest rank from the graph may expose the implementation of the course.

- Calculating the ranks by the number of routes in the graph which contains the given node, the highest ranking values are assigned to those topics that are most often included in the routes mastering different topics. Most often the following topics are included in different

learning routes: 'Overview of the course object-oriented modelling', 'Introduction into Object-Oriented Modelling', 'Unified Modelling Language – UML', 'UML possibilities in IS design', and 'UML diagrams'. Using this type of ranking calculation, higher ranks obtain the output of the system and their closest predecessors. That has been proved also in this case: the topic with the highest rank is at the top.

- Analyzing the study course topic's graph structure by the attainable number of nodes, authors can conclude that the most important topics are 'An introduction to object-oriented modelling', 'Unified Modelling Language - UML', 'UML opportunities in development of information system', 'UML diagrams', 'Structure diagram: Class diagram'. Further follow other topics.

- Assessing the results according to the elements' total rank, the most important topics are 'An introduction to object-oriented modelling', 'Unified Modelling Language - UML', 'UML diagrams', 'UML opportunities in development of information system', 'Structure diagram: The class diagram'. If from the study programme is excluded any study course and instead is introduced a new one or a new course is introduced additionally then, in order to perform correct correlation with other courses, the mutual correlation of the new course topics should be carried out against topics of other courses.

The determination of the most significant elements in graph G_2 is needed to help the lecturer make decisions about what topics include for the exam. The study course includes many topics; all cannot be included in the exam thus the determination of the most significant topics helps the lecturer to make the final decision which topics to include in the exam. When preparing the exam issues, the lecturer can assign for each exam issue points that form the evaluation, and in this case the determination of the most significant elements helps the lecturer to make the decision what points to assign to each issue, the most of the points assigning to the most significant topics. The determination of the most significant topics helps the lecturer prepare the study material because the lecturer may pay more attention to the topics which have high significance. That does not mean that rest of the topics are neglected.

For the analysis of the concept map structure there are used concepts of the topic 'Class Diagram Components'. Reviewing the summary table of the structure analysis, it can be concluded:

- Performing the structure analysis by local levels the highest rank is calculated for the concept 'Types of Links', that means that this concept has most direct links with other concepts. The next significant concepts by local levels are with equal values 'Class' and 'Cardinality', followed by 'Visibility' and with equal ranking values 'Links', 'Operations' and 'Attributes'.

- Calculating the ranks by the number of routes which contain the given concept, the highest ranking values are for the concepts 'Packages', 'Links', 'Class', 'Types of Links', and 'Operations'.

- Analyzing the structure of the study course topic graph by the number of attainable nodes, most routes are possible from the concepts 'Packages', 'Links', 'Types of Links', 'Class', and 'Cardinality'.

- Evaluating the results by the structural significance, the first most significant concepts are ranked as follows: 'Packages', 'Links', then with equal values are 'Types of Classes' and 'Class', followed by 'Cardinality'. These are five the most significant concepts. In order the learner can properly master the knowledge on topic 'Class Diagram Components', it is compulsory to master the most significant concepts in the structure. Assessing the knowledge of the learner when comparing the concept map created by the learner to the concept map defined by the lecturer, it is advisable to take into account the significance of the nodes, that means, that more points in the assessment are given if the learner has correctly identified significant concept and its correlation to other concepts than less significant concept.

In order to determine the significance of nodes for study programmes, courses or topics it is possible to use the method of structural modelling which justifies making definite decisions about changes in the study programmes, courses, and topics for both the learner and the administrative employee of the institution.

VI. CONCLUSIONS

The authors previously had to carry out research on the personalized planning of studies on graph-based framework and it is realized by the tutoring module of intelligent tutoring system. Based on described graph-based framework, SPS prototype of personalized study planning has been developed which allows the learner to design the study programme using the graph. In this article the authors describe the use of structure modelling methods to analyze personalized study planning structure based on graph framework. To carry out the qualitative analysis of the structure, ranks are calculated for its nodes. The use of structural modelling approach to analyze the structure of study programme, study courses, and concept maps is a new application of it. So far it was mainly oriented to analyze the structure of complicated systems with physically multiple elements, and technical diagnostics. In structure modelling when obtaining the ranks, the total rank is taken into account not each separately, but in this paper each rank is analyzed separately, because each has its role in the structure analysis. The authors offer also a new solution for the calculation of the structural significance values of the nodes, using 3 ranks: by local levels, by the number of the routes in the graph, and by the number of attainable nodes.

REFERENCES

[1] Rollande R., Using graph for tutor module construction of intelligent tutoring system, 4th International Conference: Information Society and Modern Business, The Role of Regional Centers in Business Development, Ventspils, Latvia, pp. 39-48, 2009.

[2] Grundspenkis J., Rollande R., Graph based framework for personalization of education process realized by the tutoring module of intelligent tutoring system, Proceeding of The International Conference «Perspectives in Business Informatics Research», 2nd International Workshop on Intelligent Educational Systems and Technology-

enhanced Learning (INTEL-EDU 2011), Riga, Latvia, pp. 216 – 225, 2011.

[3] Rollande R., Grundspenkis J., Representation of study program as a part of graph based framework for tutoring module of intelligent tutoring system, Proceedings of The Second International Conference on Digital Information Processing and Communications (ICDIPC 2012), Klaipeda, Lithuania, pp. 108-113, 2012.

[4] Rollande R., Grundspenkis J., Graph based framework and its implemented prototype for personalized study planning, The Second International Conference on E-Learning and E-Technologies in Education (ICEEE2013), Lodz, Poland, pp. 137 – 142, 2013.

[5] Rollande, R., Grundspenkis, J., Mislevics, A., The use of structural modelling methods for analysis of personalized study planning. IEEE Technically Co-Sponsored Science and Information Conference 2014 (SAI), pp. 921 - 926, London, UK, 2014.

[6] Ministry of Environmental Protection and Regional Development, Latvia 2030, 2008. Available: http://www.latvija2030.lv/upload/latvija2030_lv.pdf. [Accessed: 25.03.2013].

[7] The Council Of The European Union, Council conclusions on the modernisation of higher education, Official Journal of the European Union, pp. 36-41, 2011.

[8] The Latvian Parliament, Education Law, Latvijas Vestnesis, Riga, 1998.

[9] The Council Of The European Union, Council Conclusions on investing in education and training — a response to 'Rethinking Education:Investing in skills for better socio-economic outcomes' and the '2013 Annual Growth Survey', Official Journal of the European Union, pp. 5-8, 2013.

[10] Cabinet of Ministers, About the program «Information and communication technologies for the quality of education 2007.–2013.gadam», Latvijas Vestnesis, Latvija, 2006.

[11] Finland Ministry of Education and Culture, Education and Research 2011–2016, Government, Finland, 2012.

[12] Geer R., Sweeney T., Students' Voices about Learning with Technology, Journal of Social Sciences, vol. 8, no. 2, pp. 294, 2012.

[13] S. Kabicher and R. Motschnig-Pitrik, «Coordinating Curriculum Implementation Using Wiki-supported Graph Visualization», 9th IEEE International Conference on Advanced Learning Technologies. ICALT 2009, Riga, 2009.

[14] P. Gestwicki, «Work in progress - Curriculum visualization», Proceedings of the 38th. Annual Frontiers in Education Conference, Saratoga Springs, NY, 2008.

[15] P. Gestwicki and A. Toombs, «Poster: Design and evaluation of an interactive curriculum visualization system», VisWeek 2010, Salt Lake City, USA, 2010.

[16] R. Zucker, «ViCurriAS: a curriculum visualization tool for faculty, advisors and students», Journal of Computing Sciences in Colleges, pp. 138-145, Vol. 25 , Issue 2, 2009.

[17] T. Auvinen, «Curriculum Development Using Graphs of Learning Outcomes», First EUCEET Association Conference New Trends and Challenges in Civil Engineering Education, Patras, Greece, 2011.

[18] R. Nkambou, C. Frasson and G. Gauthier, «A new approach to ITS-curriculum and course authoring: the authoring environment», Computers & Education, Vol. 31, No. 1, pp. 105–130, august 1998.

[19] R. Nkambou, G. Gauthier and C. Frasson, «CREAM-Tools: An authoring environment for curriculum and course building in an intelligent tutoring system», CALISCE '96 Proceedings of the Third International Conference on Computer Aided Learning and Instruction in Science and Engineering, London, UK, 1996.

[20] R. Nkambou, M.-C. Frasson and C. Frasson, «Generating courses in an intelligent tutoring system», 1996.

[21] R. Nkambou, C. Frasson and G. Gauthier, «CREAM-Tools: An authoring environment for knowledge engineering in intelligent tutoring systems», In Authoring Tools for Advanced Technology Learning Environments, Springer, Netherlands, pp. 269-308, 2003.

[22] Microsoft, «.NET Framework» [Online]. Available: http://msdn.microsoft.com/en-US/vstudio/aa496123.aspx. [Accessed 02.05.2013].

[23] Microsoft, «.NET Downloads, Developer Resources & Case Studies» [Online]. Available: http://www.microsoft.com/net/. [Accessed 02.05.2013].

[24] Microsoft, «Windows Presentation Foundation» [Online]. Available: http://msdn.microsoft.com/en-us/library/vstudio/ms754130.aspx. [Accessed 02.05.2013].

[25] Microsoft, «Visual Studio LightSwitch» [Online]. Available: http://msdn.microsoft.com/en-us/vstudio/ff796201.aspx. [Accessed 02.05.2013].

[26] Microsoft, «SQL Server 2012 Express LocalDB» [Online]. Available: http://msdn.microsoft.com/en-us/library/hh510202.aspx. [Accessed 02.05.2013].

[27] Microsoft, «SQL Server Editions» [Online]. Available: http://www.microsoft.com/sqlserver/en/us/editions.aspx. [Accessed 02.05.2013].

[28] Microsoft, «Visual Studio 2012» [Online]. Available: http://msdn.microsoft.com/en-us/library/vstudio/dd831853.aspx . [Accessed 02.05.2013].

[29] Microsoft, «QuickGraph, Graph Data Structures And Algorithms for .NET» [Online]. Available: http://quickgraph.codeplex.com/. [Accessed 02.05.2013].

[30] Microsoft, «Graph#» [Online]. Available: http://graphsharp.codeplex.com/. [Accessed 02.05.2013].

[31] J. Osis, J. Grundspenkis and Z. Markovics, Complex heterogeneous systems topological modelling, Riga: Riga Technical University, 2012.

[32] Nikolaev V., Bruk V. M., Sistemotechnika: metody i prilozhenija, Leningrad: Machinostrojenie, 1985.

[33] Nečiporenko V. I., Strukturnyj analiz sistem: effektivnost'i nadežnost'., Moskva: Sovetskoe radio, 1977.

Annex 1

TABLE I.   STRUCTURE ANALYSIS SUMMARY TABLE OF THE COURSE 'OBJECT-ORIENTED MODELLING'

Title of the course topics	R^1_{LP} value	R^1_{LP} rank	R^1_{CE} value	R^1_{CE} rank	R^2_{CE} value	R^2_{CE} rank	R_{sum}	R_{tot}	N(i)	$N^1(i)$
Introduction to object-oriented modeling (OOM).	4	5	22	2	0,963	1	3	1	1	1
Unified Modeling Language - UML.	6	4	19	3	0,8148	2	5	2	0,9	0,96
UML possibilities in Information Systems (IS) design	4	5	11	4	0,7778	3	7	3	0,8	0,89
UML diagrams.	18	2	10	5	0,7407	4	9	4	0,7	0,93
Overview of the course object-oriented modelling	23	1	23	1	0,037	10	11	5	0,6	0,89
Structure diagram: Class diagram.	12	3	6	6	0,6296	5	11	5	0,6	0,85
The system dynamic model.	6	4	5	7	0,2593	6	13	6	0,5	0,81
Behavior diagram: Activity diagram.	4	5	3	8	0,1481	7	15	7	0,4	0,78
UML implementation environmt.	4	5	2	9	0,1111	8	17	8	0,3	0,74
Interaction diagram: Interaction overview diagram.	4	5	2	9	0,1111	8	17	8	0,3	0,74
IS design using object-oriented approach.	3	6	1	10	0,1111	8	18	9	0,2	0,7
Interaction diagram: Sequence diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Structure diagram: Component diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Structure diagram: Object diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Behavior diagram: State Machine diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Behavior diagram: Use - Case diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Code generation options.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Insight into the Model Driven Architecture - MDA	3	6	1	10	0,0741	9	19	10	0,1	0,67
Interaction diagram: Collaboration diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Structure diagram: Deployment diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Structure diagram: Package diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Structure diagram: Composite structure diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Interaction diagram: Timing Diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Structure diagram: Profile diagram.	3	6	1	10	0,0741	9	19	10	0,1	0,67
Basic concepts of object-oriented modeling.	2	7	1	10	0,0741	9	19	10	0,1	0,63
A brief insight into the history of the OOM.	2	7	1	10	0,0741	9	19	10	0,1	0,63
Object-oriented modeling techniques.	1	8	1	10	0,0741	9	19	10	0,1	0,59

Extended 4-Dimensional OpenGL e-book associated with electric material

Kazu-masa YAMADA

Dept. of Production Systems Engineering
Hakodate National College of Technology
Hakodate-city, Japan

Nobuaki MATSUHASHI

Dept. of Electrical and Computer Engineering
Hachinohe National College of Technology
Hachinohe-city, Japan

Abstract—The aim of this research is to develop a kind of 4 Dimensional electronic textbook (4D-Text) regarding a typical dielectrics material structure of Perovskit crystalline formations of Barium Titanate where 4D means a combined 1D direction-freely-viewing and 3D animation-ing is continuously extracting while being changed and scaled viewpoint as user chosen continuously. It is specific issue that e.g. Barium Titanium (IV) Oxide which crystallography 4D structural animation is discussed and relevantly addressing to virtual learning environments, e-learning tools, educational systems design and e-learning organizational issues. Additionally it should be an actually-expected theme that crystalline Perovskite structure with a chemical formula ABX_3 is constructed as follows; the type of anion spheres are X atoms (usually oxygens as $O[2-]$), the another type of spheres are B-atoms (a smaller metal cation, such as $Ti[4+]$) and the third type of spheres are the A-atoms (a larger metal cation, such as $Ba[2+]$). Then thermal transformation between lower temperature ferroelectorode and higher temperature paraelectorode would be discussed. Meanwhile, crystallographic pictured 4D-Text is as follows among these structures; the undistorted isometric-regular cubic, on the other hand, the symmetry lowered orthorhombic, tetragonal and trigonal in each Perovskites. It has been concluded that free OpenGL assisted 4D animation approach technique should be good way to achieve the organization for free 4D-Text e-learning using Free 3D viewing and manipulating the 4D animation data file created via MGF (MicroAVS Geometry File) approached. Additionally crystallography numerical data processing methods in Perovskit structure are to be using Freeware AWK language data-processing and extracting how to prepare the 4D with MGF.*

Therefore 4D-OpenGL Free e-Text is a possible guidance that enables studying Perovskit structure while changing a viewpoint directly to the instructor's guidance and accessing contents, transcribed by being captured screens and supplementary reference with properly alphanumeric characters and relevant information. Consequently, for the achievement of presented aim of Free 4D e-learning organizational issues, the technology of free-OpenGL is indispensable.*

Keywords—component; OpenGL; e-textbook Perovskit structure; dielectrics material; Barium Titanate

I. INTRODUCTION

It has been circulated many electronic textbook (e-textbook) particularly in medical and biomedical areas and sections. Meanwhile these e-textbook is almost static type printing e.g. pdf type. There is even superior e-textbook content for medical operations to be able to overlook the image which the internal organs specimen, but most e-textbook do not almost reach enough they do not have an ability to introduce an image, that have no advantage over to preparation image for real specimens[1,2]. In addition, it has been now still developing circulated in the e-books market field involved in areas withoutside in medical and biomedical areas.

On the other hand in hand-craft-work level, recent computational graphic technology assisted 4D digital-data were strongly supported in our e-textbook field where 4D means a combined 1D direction-freely-viewing and 3D animation-ing is continuously extracting while being changed and scaled viewpoint as user chosen continuously. The representation procedure was developed by especially free-OpenGL (Open Graphics Library) that is a multi-language application programming interface for rendering multi-dimension vector graphics, and is typically specialized to interact with a graphics processing unit to achieve hardware-accelerated rendering and with additionally interactive user-interface[3].

On the contrary meanwhile, in the material science education field, although now presenting ferroelectorode topics that is useful materials in the field of especially electronics additionally sensor-microtechnology, and higher temperature paraelectorode and lower ferroelectorode would be discussed difficultly because of the structural complexition of Perovskit structure system. But all the corners of the earth the ferroelectorode property is useful in industrial field[4].

So it is the aim of this research is to develop the e-textbook associated with Perovskit between the ferroelectorode and the paraelectorode regarding phase transition popery described and global di-electrics specification e-learning assistances with hand-craft-work level and free-software level procedures.

II. PRACTICAL PROCEDURE FOR THE WAY TO 4D REPRESENTATION

In this research firstly the totality free-software "3D AVS player" produced for only the MS-Windows platform was embraced[5]. The free-based 3D AVS player is distributed as one of the data visualization toolkit of fee-based(not free) AVS/Express series presented by Advanced Visual Systems Inc. since 1991 and CYBERNET SYSTEMS CO.,LTD. assisted with Open-GL technology and additionally the 4D digital-data supported where 4D means a combined 1D direction-freely-viewing and 3D animation-ing is continuously extracting while being changed and scaled viewpoint as user chosen continuously. Furthermore the fee-based(not free) Micro-AVS (AVS/Express function-reduced version) was also

presented by CYBERNET SYSTEMS CO.,LTD. In order to keep a limitation for free-based development environments for data processing and also to keep free-based e-textbook distribution with handcraft-works level, in presented research it has been only used such a free-based "3D AVS player" and also free-based AWK which is an free-based interpreted programming language designed for text processing and typically used as a data extraction and reporting tool within Linux OS, Macintosh OS and optional-installed MS Windows OS.

In the Figure 1, an overall process illustrated flowchart has been shown that is the first hand-made prepared the multi-byte ASCII based static file that is MicroAVS Geometry File so called MGF, afterwards final using AWK data-processing automatically being to extract into the animation type MGF ASCII based movie file (i.e. the goal animation.mgf was finally produced).

Additionally in this figure, using step sequence coordinated with AWK, the #0.mgf, #1.mgf, …#n.mgf, each one is all static type (not animation type) MicroAVS Geometry File with OpenGL-technology-enabled 3D numerical data with different point ways enabled viewing. Addition all gathered MGFs are compressed with OpenGL-technology-enabled 4D , then single animation MGF would be produced assisted with handmade AWK script.
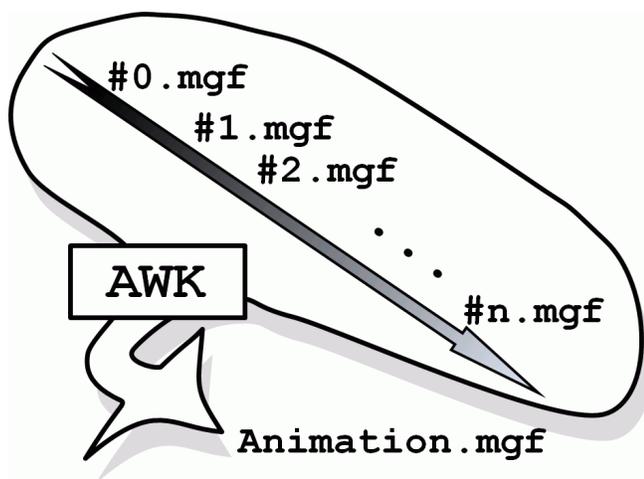
Fig.1. Schematic illustration for step sequence coordinated with AWK , which is an interpreted programming language designed for text processing and typically used as a data extraction and reporting tool. The #0.mgf, #1.mgf, …#n.mgf, each one is all static type (not animation type) MicroAVS Geometry File with OpenGL-technology-enabled 3D numerical data with different point ways enabled viewing. Addition all gathered MGFs are compressed with OpenGL-technology-enabled 4D , then single animation MGF would be produced assisted with handmade AWK script.

In unit of angstrom order of Fig. 2, listed and schematic illustration has been shown, on the contrary in Fig.1 for the #0.mgf, #1.mgf, …#n.mgf, which is static type (not animation type) MicroAVS Geometry File with OpenGL-technology-enhanced 3D numerical data with different point ways enabled viewing. Taking objection, in Fig. 2, left figure and right one are the #0.mgf i.e. 00static.mgf and the #1.mgf i.e. 01static.mgf, respectively. Afterward, gathered them of #0.mgf, #1.mgf, …#n.mgf MGFs are compressed with OpenGL-technology-enabled 4D animation MGF file.

Additionally in Fig. 2, contents which were common to both, these list means as below:

- # Micro AVS Geom:2.00, the 1st line, the common header for Micro AVS Geometry ASCII file.

- sphere, the 2nd line must be concrete order.

- BaTiO3_for_Ba, the 3rd line, the arbitrary character strings.

- color, the 4th line, the vertex data type.

- The 5th line number, the component number.

- In presented case, it should be described basic element type "sphere" in the second line, and the repetition.

- It should be described any name in the third line and the repetition.

- It should be described vertex data type in the fourth line and the repetition.

- sphere, it should be described definition the sphere which was appointed in central coordinate level, radius, having color or not, and the repetition.

- The vertex data type has "vertex" and "color", and the repetition.

- label, in the periodic domain, the 1st line must be concrete order, and the repetition.

- atom_position_label, in the periodic domain, the 2nd line must be the arbitrary character strings, and the repetition.

- color, in the periodic domain, the 3rd line must be the concrete order, and the repetition.

- Courier 0.05 1 0 0 1 0.0 0.0 0.1, the 4th line number, in the periodic domain, regarding the label, font-type, size, style1, style2, title, reference-position and off-set.

- font-type: Available font is the following; Courier, Helvetica, New Century, Times, Charter, Symbol, Roman, Script, Mathematics. You can appoint here from these font-type.

- size: definition of font size.

- style1: definition of font thickness, 0=nomal, 1=bold.

- style2: definition of font italic, 0=nomal, 1=italic.

- title: definition of title fixture, 0=nomal, 1=fixed.

- reference-position: definition of reference-position, 0=alignleft, 1=centering 2=alignright.

- off-set size: definition of off-set, X Y Z.

- Caution font-type: label character must be less than 300. It should be surround with double quotation. Any metacharacter with special meaning may be quoted by preceding it with a backslash.

- "4", the 5th line number, the component number of lavel.

- The 6th line and the later, in the periodic domain, regarding the lavel, position(X, Y, Z), color(R, G, B) and the lavel of the arbitrary character strings with double-quotes.

- column, in the periodic domain, the 1st line must be concrete order, and the repetition.

- cylinder_2, in the periodic domain, the 2nd line must be the arbitrary character strings, and the repetition.

- dvertex_and_color, in the periodic domain, the 3nd line must be the concrete order, and the repetition.

- The 4th line number, in the periodic domain, the number of partitions of the column, and the repetition.

- The 5th line number, in the periodic domain, regarding the column, direction(X, Y, Z), radius, altitude and color(R, G, B), and the repetition.

- disjoint line, in the periodic domain, the 1st line must be concrete order, and the repetition.

- x_y_z_axis-drawing, in the periodic domain, the 2nd line must be the arbitrary character strings, and the repetition.

- color, in the periodic domain, the 3rd line must be the concrete order, and the repetition.

- "6", the 4th line number, the component number of vertex data for disjoint line.

- Caution disjoint line: the component number "6" of vertex data for disjoint line means 3 disjoint lines.

- The 5th line and the later on disjoint line part, in the periodic domain, regarding the vertex, position(X, Y, Z), and color(R, G, B).

- Caution the color for disjoint line: disjoint line have 2 point of vertex with deferent colors, it should be gradation color.

- The final # x y z radius R G B is the simple comment line, it should be able to add some comment line only in final domain, it should be added # to a head of a line.

In Fig 3, it has been shown a script-formed illustration for typical language rule optimized for standard AWK which is an interpreted programming language designed for text processing and typically used as a data extraction and reporting tool within Linux OS, Macintosh OS and opptional-installed MS Windows OS. Afterwards, the "result01" directory including in fig.1 of #0.mgf, #1.mgf, …#n.mgf, i.e. in fig 3 regarding practically for 00static.mgf and 01static.mgf were stored. Addition all gathered these MGFs are compressed and re-produced single 4D-type animation type MicroAVS Geometry File.

Afterward in Fig 4. for only the MS-Windows platform, script-formed illustration for typical batch-file for using gawk MS-DOS command-line which is an interpreted control command programming for batch sequence. The "multidata2movie.awk.bat" is the presented batch-filename for producing a 4D-type animation redirect to "movie.mgf".

Above these files are summarized and typical examples are shown in Fig 5. for only the MS-Windows platform, list-formed illustration for typical working directory structure by using version 3.2.5 MBCS (Multiple Byte Character Set) gawk.exe which is an interpreted programming language designed for Multiple Byte text processing tool. In summary, the "result01" is the static type MicroAVS Geometry Files' Directory for in fig.1 of #0.mgf, #1.mgf, …#n.mgf, i.e. regarding in fig 2 for 00static.mgf and 01static.mgf (each one is 3D but not animation type) are all stored in it. Addition all gathered MGFs are compressed and produced a 4D type "animation" into presented working directory. In the meantime the animation MGF is produced and assisted with gawk.exe using "multidata2movie.awk" script, the "multidata2 movie.awk.bat" in fig 4 is the batch file for finally producing a 4D-type animation "movie.mgf".

III. RESULTS AND DISCUSSION

For typical result of a 4D-type animation, it would be used via Free-ware (no fee, but registration requested [6]) "3D AVS player" assisted by OpenGL technology for 4D moving observation according the MGF ASCII formatted file. Hence in Fig. 6 the 4D-type animation "movie.mgf" snapshot using "3D AVS player" has been shown, where is a 64bit-type frame shot and with "RNC" meaning "Reset, Normalize, Center" (functional relation Ctrl+D), then with "B" meaning Bottom (Ctrl+O), to obtain a more accurate observation the movie, select the "perspective view" (Ctrl+P) of "non parallel perspective mode", play mode "cycle" and "play" (Shift+R).

Left box — #0.mgf, i.e. 00static.mgf

```
# Micro AVS Geom:2.00
sphere
BaTiO3_for_Ba
color
8
  .0    .0   .0     1.35    0.2   0.2   0.6
3.992  .0   .0     1.35    0.2   0.2   0.6
  .0  3.992 .0     1.35    0.2   0.2   0.6
3.992 3.992 .0     1.35    0.2   0.2   0.6

  .0    .0  4.038   1.35   0.2   0.2   0.6
3.992  .0  4.038   1.35   0.2   0.2   0.6
  .0  3.992 4.038  1.35   0.2   0.2   0.6
3.992 3.992 4.038  1.35   0.2   0.2   0.6
sphere
BaTiO3_for_O
color
6
1.996  1.996  .0     1.40   0.9   0.9   0.9
1.996   .0   2.019  1.40   0.9   0.9   0.9
  .0  1.996  2.019  1.40   0.9   0.9   0.9
1.996  3.992 2.019  1.40   0.9   0.9   0.9
3.992  1.996 2.019  1.40   0.9   0.9   0.9
1.996  1.996 4.038  1.40   0.9   0.9   0.9
sphere
BaTiO3_for_Ti
color
1
1.996  1.996  2.019  0.76  0.98   0   0
label
atom_position_label
color
Courier 0.05 1 0      0     1     0.0 0.0 0.1
4
  .0    .0   .0  0.8 0.8 0.8 ".0    .0    .0"
3.992  .0   .0  0.8 0.8 0.8 "3.992 .0   .0"
  .0  3.992 .0  0.8 0.8 0.8 ".0  3.992 .0"
  .0    .0 4.038 0.8 0.8 0.8 ".0    .0  4.038"
column
cylinder_2
dvertex_and_color
16
12
.0 .0 .0    .0     .0    4.038    0.06    .6 .0 .0
.0 .0 .0    .0   3.992    .0      0.06    .6 .0 .0
.0 .0 .0  3.992   .0      .0      0.06    .6 .0 .0
3.992 3.992 4.038   .0  3.992  4.038  0.06  .6 .0 .0
3.992 3.992 4.038  3.992  .0   4.038  0.06  .6 .0 .0
3.992 3.992 4.038  3.992 3.992  .0    0.06  .6 .0 .0
  .0    .0  4.038  .0   3.992  4.038  0.06  .6 .0 .0
  .0    .0  4.038 3.992  .0    4.038  0.06  .6 .0 .0
  .0  3.992   .0    .0   3.992  4.038  0.06  .6 .0 .0
  .0  3.992   .0  3.992 3.992   .0    0.06  .6 .0 .0
3.992  .0    .0  3.992 3.992   .0    0.06  .6 .0 .0
3.992  .0    .0  3.992  .0    4.038  0.06  .6 .0 .0
label
title-label
color
Courier 0.1 1 0       1     1     0.0 0.0 0.1
2
0 -1.00 0.0  0.0 0.0 0.8 "(C)YAMA@hakodate-ct.ac.jp"
0 -0.95 0.0  0.8 0.0 0.0 "BaTiO3 structure"
label
axis_and_coordinate_origin
color
Courier 0.05 1 0      0     1     0.0 0.0 0.1
3
10 0 0  0.8 0.8 0.8 "X-axis"
 0 10 0  0.8 0.8 0.8 "Y-axis"
 0 0 10  0.8 0.8 0.8 "Z-axis"
disjoint line
x_y_z_axis-drawing
color
6
 0 0 0   0.8 0.8 0.8
10 0 0   0.8 0.8 0.8
 0 0 0   0.8 0.8 0.8
 0 10 0  0.8 0.8 0.8
 0 0 0   0.8 0.8 0.8
 0 0 10  0.8 0.8 0.8
# x y z radius R G B
```

Right box — #1.mgf, i.e. 01static.mgf

```
# Micro AVS Geom:2.00
sphere
BaTiO3_for_Ba
color
8
  .0    .0   1.0    1.35    0.2   0.2   0.6
3.992  .0   1.0    1.35    0.2   0.2   0.6
  .0  3.992 1.0    1.35    0.2   0.2   0.6
3.992 3.992 1.0    1.35    0.2   0.2   0.6

  .0    .0  5.038   1.35   0.2   0.2   0.6
3.992  .0  5.038   1.35   0.2   0.2   0.6
  .0  3.992 5.038  1.35   0.2   0.2   0.6
3.992 3.992 5.038  1.35   0.2   0.2   0.6
sphere
BaTiO3_for_O
color
6
1.996  1.996  .0     1.40   0.9   0.9   0.9
1.996   .0   2.019  1.40   0.9   0.9   0.9
  .0  1.996  2.019  1.40   0.9   0.9   0.9
1.996  3.992 2.019  1.40   0.9   0.9   0.9
3.992  1.996 2.019  1.40   0.9   0.9   0.9
1.996  1.996 4.038  1.40   0.9   0.9   0.9
sphere
BaTiO3_for_Ti
color
1
1.996  1.996  3.019  0.76  0.98   0   0
label
atom_position_label
color
Courier 0.05 1 0      0     1     0.0 0.0 0.1
4
  .0    .0   .0  0.8 0.8 0.8 ".0    .0    .0"
3.992  .0   .0  0.8 0.8 0.8 "3.992 .0   .0"
  .0  3.992 .0  0.8 0.8 0.8 ".0  3.992 .0"
  .0    .0 4.038 0.8 0.8 0.8 ".0    .0  4.038"
column
cylinder_2
dvertex_and_color
16
12
0 .0 .0    .0     .0    4.038    0.06    .6 .0 .0
.0 .0 .0    .0   3.992    .0      0.06    .6 .0 .0
.0 .0 .0  3.992   .0      .0      0.06    .6 .0 .0
3.992 3.992 4.038   .0  3.992  4.038  0.06  .6 .0 .0
3.992 3.992 4.038  3.992  .0   4.038  0.06  .6 .0 .0
3.992 3.992 4.038  3.992 3.992  .0    0.06  .6 .0 .0
  .0    .0  4.038  .0   3.992  4.038  0.06  .6 .0 .0
  .0    .0  4.038 3.992  .0    4.038  0.06  .6 .0 .0
  .0  3.992   .0    .0   3.992  4.038  0.06  .6 .0 .0
  .0  3.992   .0  3.992 3.992   .0    0.06  .6 .0 .0
3.992  .0    .0  3.992 3.992   .0    0.06  .6 .0 .0
3.992  .0    .0  3.992  .0    4.038  0.06  .6 .0 .0
label
title-label
color
Courier 0.1 1 0       1     1     0.0 0.0 0.1
2
0 -1.00 0.0  0.0 0.0 0.8 "(C)YAMA@hakodate-ct.ac.jp"
0 -0.95 0.0  0.8 0.0 0.0 "BaTiO3 structure"
label
axis_and_coordinate_origin
color
Courier 0.05 1 0      0     1     0.0 0.0 0.1
3
10 0 0  0.8 0.8 0.8 "X-axis"
 0 10 0  0.8 0.8 0.8 "Y-axis"
 0 0 10  0.8 0.8 0.8 "Z-axis"
disjoint line
x_y_z_axis-drawing
color
6
 0 0 0   0.8 0.8 0.8
10 0 0   0.8 0.8 0.8
 0 0 0   0.8 0.8 0.8
 0 10 0  0.8 0.8 0.8
 0 0 0   0.8 0.8 0.8
 0 0 10  0.8 0.8 0.8
# x y z radius R G B
```


Fig.2. In unit of angstrom order, listed and schematic illustration for the #0.mgf, #1.mgf, …#n.mgf, which is static type (not animation type) MicroAVS Geometry File with OpenGL-technology- enhanced 3D numerical data with different point ways enabled viewing. Left figure and right one are the #0.mgf i.e. 00static.mgf and the #1.mgf i.e. 01static.mgf, respectively. Afterward, gathered them of #0.mgf, #1.mgf, …#n.mgf MGFs are compressed with OpenGL-technology-enabled 4D animation MGF file. The underline value are the differenc between the left and the right.

```
# AWK script: multidata2movie.awk "(C)YAMA@hakodate-ct.ac.jp"
BEGIN{

pagination=2;

prefix_dat_file_name="result01/";
suffix_dat_file_name="static.mgf";

print "# Micro AVS Geom:2.00"
print "#multidata2movie"
print pagination;

#-------------------------------------------

for(page=1;page<=pagination;page++){

print "step" page;    #page1,2,3,,,,,,,,,pagination
print ""
print ""
print ""

#-------------------------------------------
#filename has a number which has one numeral less than the pagination, then
applied -1.

if(page<=10){
dat_file_name= prefix_dat_file_name  0  page-1  suffix_dat_file_name;
#e.g. result01/00static.mgf, result01/01static.mgf,,,,result01/09static.mgf
}else{
dat_file_name= prefix_dat_file_name  page-1  suffix_dat_file_name;
#e.g. result01/10static.mgf, result01/11static.mgf,,,,result01/99static.mgf
};

print " dat_file_name =" dat_file_name > "/dev/stderr"
 #command line massage output
#-------------------------------------------

getline < dat_file_name; #the "Micro AVS Geom:2.00" line must be skipped.

while(getline < dat_file_name >0){print;}
 #every line should be outputted(printed).

print ""
print ""
print ""
print "label";
print "title-label" # discretionary one word
print "color" # vertex or color
print "Courier  0.1  1  0      1  1      0.0  0.0  0.1"
# font-type  size  style1  style2  title  reference-position  off-set
# font-type: Available font is the following; Courier, Helvetica, New Century, Times,
Charter, Symbol, Roman, Script, Mathematics.  You can appoint here from these
font-type.
# size: definition of font size.
# style1: definition of font thickness, 0=nomal, 1=bold.
# style2: definition of font italic, 0=nomal, 1=italic.
# title: definition of title fixture, 0=nomal, 1=fixed.
# reference-position: definition of reference-position, 0=alignleft, 1=centering
2=alignright.
# off-set size: definition of off-set, X Y Z.
#Caution: label character must be less than 300.  It should be surround with double
quotation. Any metacharacter with special meaning may be quoted by preceding it
with a backslash.

print "1" # number of label

printf ("0 -0.85  0.0   0.8  0  0.8  \"page=%2d, movie data\"", page);
# X Y Z R G B label

print ""
print ""
print ""

} #end of for(page=1;page<=pagination;page){
} #end of BEGIN
```

Fig.3. Script-formed illustration for typical optimized standard AWK which is an interpreted programming language designed for text processing and typically used as a data extraction and reporting tool within opptional-installed MS Windows OS and so on. Afterwards, the "result01" Directory including in Fig.1 of   #0.mgf, #1.mgf, …#n.mgf, i.e. in fig 3 regarding practically for 00static.mgf and 01static.mgf were stored.   Addition all gathered these MGFs are compressed and re-produced single 4D-type animation type MicroAVS Geometry File.

```
rem   batch file,
rem   written for only the MS-Windows platform: multidata2movie.awk.bat
gawk -f multidata2movie.awk > movie.mgf
```

Fig.4. For only the MS-Windows platform, script-formed illustration for typical batch-file for using gawk MS-DOS command-line which is an interpreted control command programming for batch sequence. The "multidata2movie.awk.bat" is the presented batch-filename for producing a 4D-type animation redirect to "movie.mgf".

```
Structure of working Directory by using gawk3.1.5-mbcs-win32-20051223:
2014/01/23  21:00   <DIR>          .
2014/01/23  21:00   <DIR>          ..
2014/01/23  22:00   <DIR>          result01
2005/12/30  00:24         446,464  gawk.exe
2014/01/23  23:00           1,534  multidata2movie.awk
2014/01/23  23:30             134  multidata2movie.awk.bat
```

Fig.5. For only the MS-Windows platform, list-formed illustration for typical working Directory structure by using ver 3.2.5 MBCS (Multiple Byte Character Set) gawk.exe which is an interpreted programming language designed for Multiple Byte text processing tool. Afterwards, the "result01" is the static type MicroAVS Geometry Files' Directory for in fig.1 of #0.mgf, #1.mgf, …#n.mgf, i.e. regarding in fig 2 for 00static.mgf and 01static.mgf (each one is 3D but not animation type) are all stored in it.   Addition all gathered MGFs are compressed and produced a 4D type "animation" into presented working Directory. In the meantime the animation MGF is produced and assisted with gawk.exe using "multidata2movie.awk" script, the "multidata2 movie.awk.bat" in fig 4 is the batch file for finally producing a 4D-type animation "movie.mgf".
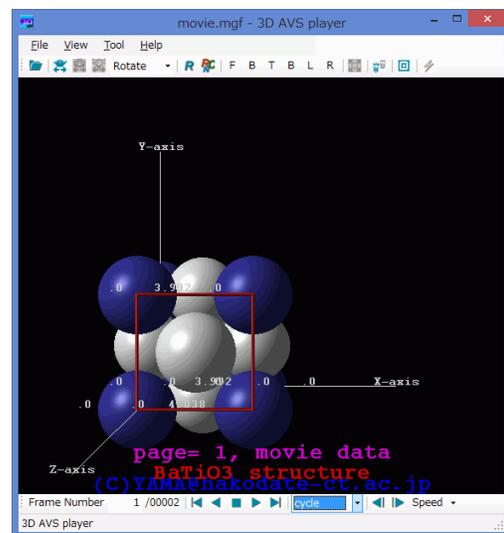

Fig.6. 4D-type animation "movie.mgf" snapshot using "3D AVS player", where a 64bit-type frame shot and with "RNC (functional relation Ctrl+D)". Further to obtain a more accurate observation the movie, select the "perspective view(Ctrl+P)", play mode "cycle" and "play(Shift+R)".

Consequence in Fig. 6, it has been obtained a 4 Dimensional electronic textbook (4D-Text) regarding a typical dielectrics material structure of Perovskit crystalline formations of Barium Titanate where 4D means a combined 1D direction-freely-viewing and 3D animation-ing is continuously extracting while being changed and scaled viewpoint as user chosen continuously. In Barium Titanium (IV of coordination number) Oxide as $BaTiO_3$, it is specific issue and would be discussed with thermal transition phenomenon between the para/ferro electorode that crystallography 4D structural animation addressing relevantly to virtual learning environments, e-learning tools, educational systems design and e-learning organization.

In Fig. 7, schematic illustration in unit of angstrom order, it has been shown two kind of snapshot among the para/ferro electorode transformation where left one of page 1 , right one of page 2, higher temperature structure of para-electorde shown as home position Ba (middle size of sphere), Oxide (large size of sphere) and Ti (small size one and hidden in Fig. 7), lower temperature structure of ferro-electorde shown as shifted position towards z-axis (i.e. c-axis in crystallography) direction with Ba and Ti(hidden), meanwhile as home position with Oxide, respectively, where in a same 4D-type animation "movie.mgf" using "3D AVS player", where to obtain a more accurate observation the movie, push "RNC (functional relation Ctrl+D)", then, push "B (Ctrl+O)" meaning Bottom, further select the "perspective view(Ctrl+P)".

Caution in Fig. 7 and throughout the presented research, there is an assumption to distinct the transformation process between the page 1 and 2, they have an assumption in the following conditional clause to perform validation on the incoming page: among in the para/ferro electorode transformation,1.0 angstrom order distance applied to the Ba (middle size of sphere) and Ti (small size one and hidden in Fig. 7), on the other hand Oxide (largest size of sphere) treatment is not indicated for this case since their polarization.

In general, it should be an actually-expected theme that crystalline Perovskite structure with a chemical formula ABX3 is constructed as follows; the type of anion spheres are X atoms (usually oxygens as O^{2-}), the another type of spheres are B-atoms (a smaller metal cation, such as Ti^{4+}) and the third type of spheres are the A-atoms (a larger metal cation, such as Ba^{2+}). Then thermal transformation between lower temperature ferroelectorode and higher temperature paraelectorode would be appear causing of their polarization with deformation, flexure and stress regarding under the Currie temperature environment. The phenomenon of stress induced would be able to refer the magneto-optical recording giant perpendicular magnetic anisotropy with Tb-Fe amorphous alloys [7, 8].

Meanwhile, crystallographic pictured 4D-Text is as follows among these structures; the undistorted isometric-regular cubic, on the other hand, the symmetry lowered orthorhombic, tetragonal and trigonal in each Perovskites.

Besides presented L. Pauling's ionic radius for BaTiO3 [9] are shown in Table I and also presented single crystalline data for pure BaTiO3 [10] are shown in Table II. It has been obtained a characteristic performance in presented 4D movie, distance and specific character for overlap between nearest neighbor ionics that can be hardly identified are recognized. Meanwhile according in these value in the tables, it is difficult to distinct the nearest neighbor ionics overlap specific character.

TABLE.I.   PRESENTED L.PAULING'S IONIC RADIUS FOR BATIO3 [9] .

Unit: angstrom order $(10^{-10}$ m)	Presented ionic radius for $BaTiO_3$		
	Ti^{4+}	Ba^{2+}	O^{2-}
Ionic radius	0.76	1.35	1.40

TABLE.II.   PRESENTED SINGLE CRYSTALLINE DATA FOR PURE BATIO$_3$ [10] .

Unit: angstrom order $(10^{-10}$ m)	Presented single crystalline data for $BaTiO_3$		
	a	*b*	*c*
Lattice constant	3.992	3.992	4.038
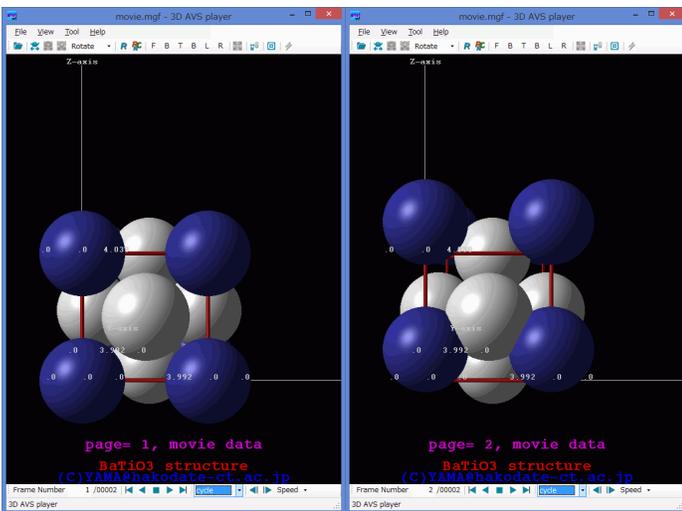
Fig.7.   Two kind of page snapshots where left one is page 1 and right one is page 2 in a same 4D-type animation "movie.mgf" using "3D AVS player", where to obtain a more accurate observation the movie, push "RNC (functional relation Ctrl+D)", then, push "B (Ctrl+O)" meaning Bottom, further select the "perspective view(Ctrl+P)".
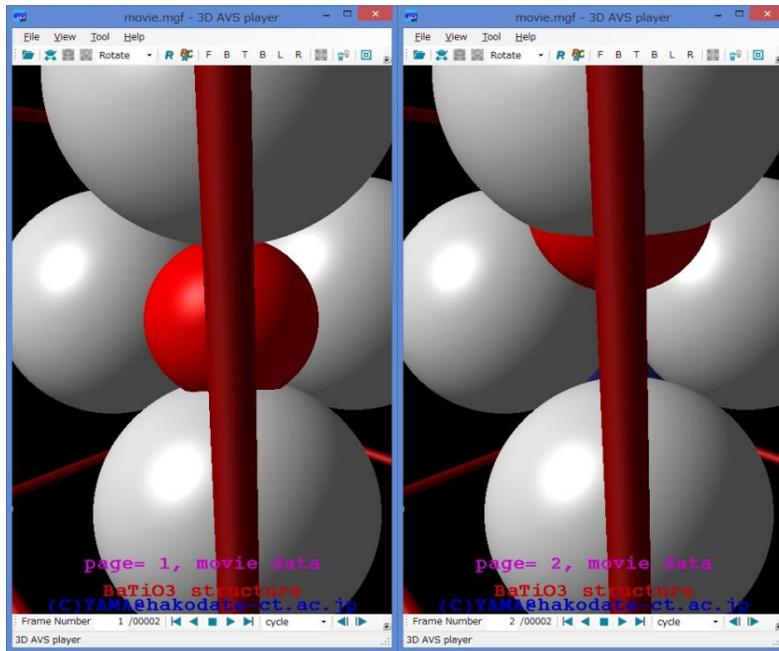
Fig.8. Zooming the two kind of page snapshots where left one is page 1 and right one is page 2 in a same 4D-type animation "movie.mgf" using "3D AVS player", where to obtain a more accurate observation the movie with "perspective view(Ctrl+P)" mode. Center position the Ti atom, it is ease to recognize the diffusinon between the structre of para/ferro transformation.
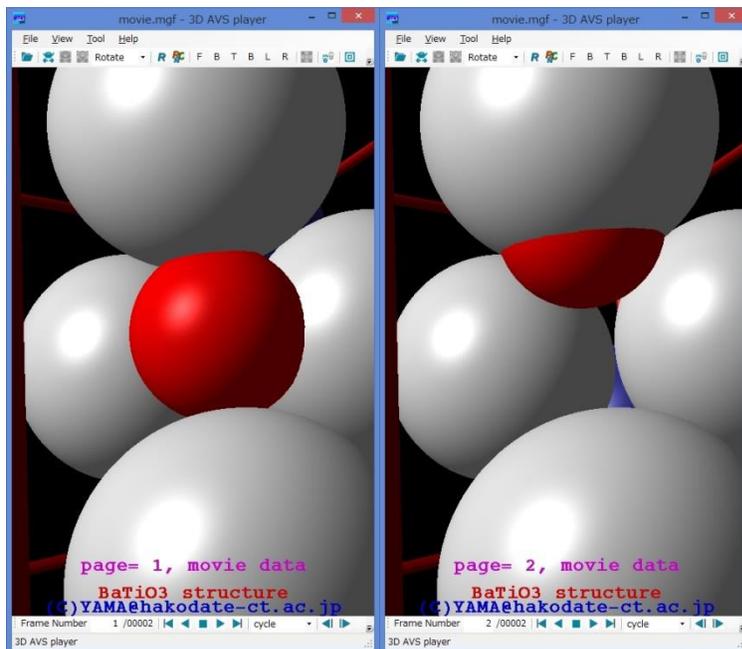
Fig.9. Since the Fig. 8 positions, extra zooming the two kind of page snapshots where left one is page 1 and right one is page 2 in a same 4D-type animation "movie.mgf" using "3D AVS player", where to obtain a more accurate observation the movie with also the "perspective view(Ctrl+P)" mode. Further the center position, also the Ti atom, it is ease to recognize the diffusinon between the structre of para/ferro transformation.

Additionally moreover, by use of the zooming operation with perspective viewing shown in Fig 8 and 9, although there is an assumption to distinct the transformation process between the higher and lower temperature of lattice transformation, i.e. among in the para/ferro electorode transformation, 1.0 angstrom order distance applied to the Ba and Ti, but Oxide treatment is not indicated for this case, it has been good method to recognize the transformation even onto the diffusion environment around hardly identified Ti atom in the place where it is easy to hide in are recognized.

It has been concluded that free OpenGL assisted 4D animation approach technique should be good way to achieve the organization for free 4D-Text e-learning using Free 3D viewing and manipulating the 4D animation data file created via MGF (MicroAVS Geometry File) approached. Additionally crystallography numerical data processing methods in Perovskit structure are to be using Freeware AWK language data-processing and extracting how to prepare the 4D with MGF.

Therefore 4D-OpenGL Free e-Text is a possible guidance that enables studying Perovskit structure while changing a viewpoint directly to the instructor's guidance and accessing contents, transcribed by being captured screens and supplementary reference with properly alphanumeric characters and relevant information. Consequently, for the achievement of presented aim of Free 4D e-learning organizational issues, the technology of free-OpenGL is indispensable.

Summarized appendix is going to be shown in Fig. 10, a almost final purchase product of the 4D "movie.mgf" is shown through a purchase procedure via awk script of Fig 3 and with the seed of static 3D MGFs of Fig 2.

IV. ADVANCED DOCUMENTATION DESCRIPTION[11]

For documentations in the OpenGL description, it would be complicated processes in general. However in MGF formatted, it will be more easy to describe only with numerical digits. We would like to show the typical rule for MGF formatted lettering as follows grammar.

The letter [11]

In generation letters. It will be defined coordinate of axes, having color / or not, a font, size, a style (bold-faces, etc) to display a letter, the character string that appointed the title letter.

label:

It is one of the types to generate a letter. It will be appointed one point of coordinate displaying any character strings.

A Grammar for Letter Descriptions Format [11]:

Micro AVS Geom: 2.10

label

Element name

vertex data type (vertex or color)

font size style-1 style-2 title base-line position offset

label number (n)

X(1) Y(1) Z(1) (R(1) G(1) B(1)) "The descriptions"

.

▪ (only the label numbers (n) repeats themselves)

.

X(n) Y(n) Z(n) (R(n) G(n) B(n)) "The descriptions"

In the second line, it will be described imperative construct type label.

In the third line, it will be described any name.

In the fourth line, it will be described vertex data type.

vertex: coordinate of axes (X,Y,Z) of each vertex

color: coordinate of axes (X,Y,Z) of each vertex, color (R,G,B)

The fifth line appoints the format of the letter.

Font: Definition font style.

Size: Definition font size.

Style-1: Definition of the letter thickness (0=standard, 1=bold-faced)

Style-2: Definition of the style italic type letters (0=standard, 1=italic type)

Title: Definition of the title (0=standard, 1=title)

Standard position: Definition of the letter position (0=Align left, 1=Align Central, 2=Align right)

Offset: Quantity of displacement from the offset standard position (X,Y,Z)

In the sixth line, it will be defined the number of the labels. It was read letter(s) which defined numericals.

In seventh line and the later, It will be described coordinate of axes (X,Y,Z), color (R,G,B) depending on vertex data type.

If the title is 1, then displayed letter was fully fixed, and the mouse operation cannot be under operation.

In the indication window, the Offset, and if the title is 1, it was the coordinate system to assume as (1,1,1) for the downstage right corner, meanwhile to assume as (-1,-1,-1) for the upstage left corner.

It will be changed by the aspect ratio of the indication window actually. Additionally the Z value influences stacking with the objects.

►A sample 1, Title letter (title2.mgf) [11]

Micro AVS Geom: 2.10

label

sample1

Color

Courier 0.2 1 0 1 1   0.0 0.0 0.1

1

0.0 0.2 0.0   0.0 0.8 0.0 "Test\" avs \"Label"

Attention:

Vertex data type does not describe RGB at the time of vertex. In this version, the Japanese font is not supported.

An available font is the following: Courier, Helvetica, New Century, Times, Charter, Symbol, Roman, Script, Mathematics

Attention:

It should be surrounded the character string with "".

When It will be used " in character string, would you put the \ (backslash character) just before them. Also letter string allows up to 300 characters.

► A sample 2, No color; a standard letter (label.mgf) [11]

Micro AVS Geom: 2.10

label

sample2

vertex

Helvetica 0.1 0 1 0 1   0.0 0.0 0.0

5

 0.0 0.0 0.0 "center"

 0.2 0.0 0.0 "right"

-0.2 0.0 0.0 "left"

 0.0 0.2 0.0 "top"

 0.0 -0.2 0.0 "bottom"

V.   SUMMARY

4D means a combined 1D direction-freely-viewing and 3D animation-ing continuously extracting while being changed and scaled viewpoint as user chosen continuously. whereby it is useful for e-text and e-learning materials regarding the use of the static and dynamic type MicroAVS Geometry Files' tools. Additionally 4D-Textbook regarding a typical dielectrics material structure of Perovskit crystalline formations of Barium Titanate where 4D "movie.mgf" are easy to be produced by using AWK numerical data processing. And also non-fee 4D viewing good tool was ease to supplied environmentally through the OpenGL assisted computational technology. It can be recognized the effectivity of the 4D e-text for e-learning Perovskit crystalline formations of $BaTiO_3$.

REFERENCES

[1] BBC Online News, "Can stem cells heal broken hearts?," http://bbc.in/1ffFOUJ , 20th February 2014.

[2] Barbara Janson Cohen, "Memmler's The Human Body in Health and Disease," Lippincott Williams & Wilkins, 1st January 2012.

[3] The OpenGL Graphics System: A Specification (Version 4.0 Core Profile) Mark Segal and Kurt Akeley, http://www.opengl.org /registry/doc/glspec40.core.20100311.pdf, March 2010.

[4] Electrode Systems, Ferro Corporation, http://www.ferro.com/non-cms/ems/epm/content/docs/e1192.pdf, January 2011.

[5] Gawk MBCS (Multiple Byte Character Set), http://www.vector.co.jp/ soft/dl/win95/util/se376460.html, December 2005.

[6] Free 3D Viewer, "3D AVS Player version 3.2", 3DAVSplayer_v32_32bit.zip and 3DAVSplayer_v32_64bit.zip, http://www.cybernet.co.jp/avs/english/avsplayer.html, Copyright 2014 CYBERNET SYSTEMS CO.,LTD.

[7] K. Yamada, A. Fujita, M. Ohta and K. Fukamichi, "Disappearance of Perpendicular Magnetic Anisotropy in Amorphous Tb-Fe Alloys," Journal of Magnetism and Magnetic Materials. Vol. 239, pp. 412-414, February 2002.

[8] Motoki Ohta, Kazumasa Yamada, Yoshihiko Satake, Asaya Fujita and Kazuaki Fukamichi, "Origin of Perpendicular Magnetic Anisotropy in Tb-Fe Amorphous Alloy", Materials Transactions, Vol.44 No.12 pp. 2605-2610, December 2003.

[9] L. Pauling, The Nature of the Chemical Bond, 3rd Edn., Cornell University Press, Ithaca, N. Y. ,1960.

[10] Lattice constants for single crystalline $BaTiO_3$ cited from the single crystalline database of Techno Chemics, Inc. (crystals.jp/eindex.html), http://crystals.jp/BaTiO3.html, March 2014.

[11] Usermanual11.pdf, in Japanese, Produced by CYBERNET SYSTEMS CO.,LTD., "Users manual for MicrAVS including the MGF format, etc" http://www.cybernet.co.jp/avs/download/microavs_tech.html , 14th October 2014.

```
# Micro AVS Geom:2.00
#multidata2movie
2
step1


sphere
BaTiO3_for_Ba
color
8
  .0    .0   .0     1.35     0.2     0.2     0.6
3.992  .0   .0     1.35     0.2     0.2     0.6
  .0  3.992 .0     1.35     0.2     0.2     0.6
3.992 3.992 .0     1.35     0.2     0.2     0.6

  .0    .0  4.038  1.35     0.2     0.2     0.6
3.992  .0  4.038   1.35     0.2     0.2     0.6
  .0  3.992 4.038  1.35     0.2     0.2     0.6
3.992 3.992 4.038  1.35     0.2     0.2     0.6

sphere
BaTiO3_for_O
color
6
 1.996  1.996  .0    1.40   0.9    0.9    0.9
 1.996   .0   2.019  1.40   0.9    0.9    0.9
  .0   1.996  2.019  1.40   0.9    0.9    0.9
 1.996 3.992  2.019  1.40   0.9    0.9    0.9
 3.992 1.996  2.019  1.40   0.9    0.9    0.9
 1.996 1.996  4.038  1.40   0.9    0.9    0.9

sphere
BaTiO3_for_Ti
color
1
 1.996  1.996  2.019  0.76   0.98    0     0

label
atom_position_label
color
Courier 0.05 1 0        0       1        0.0 0.0 0.1
4
  .0    .0   .0  0.8 0.8 0.8 " .0   .0   .0"
3.992  .0   .0  0.8 0.8 0.8 "3.992  .0   .0"
  .0  3.992 .0  0.8 0.8 0.8 " .0  3.992 .0"
  .0    .0  4.038 0.8 0.8 0.8 " .0   .0  4.038"

column
cylinder_2
dvertex_and_color
16
12
.0 .0 .0   .0    .0   4.038   0.06    .6 .0 .0
.0 .0 .0   .0  3.992  .0      0.06    .6 .0 .0
.0 .0 .0  3.992  .0   .0      0.06    .6 .0 .0

3.992 3.992 4.038   .0  3.992 4.038  0.06  .6 .0 .0
3.992 3.992 4.038  3.992  .0   4.038  0.06  .6 .0 .0
3.992 3.992 4.038  3.992 3.992  .0    0.06  .6 .0 .0

  .0    .0  4.038 .0 3.992 4.038  0.06  .6 .0 .0
  .0    .0  4.038 3.992 .0 4.038  0.06  .6 .0 .0
  .0  3.992  .0  .0 3.992 4.038  0.06  .6 .0 .0
  .0  3.992  .0  3.992 3.992 .0   0.06  .6 .0 .0
3.992  .0    .0  3.992 3.992 .0   0.06  .6 .0 .0
3.992  .0    .0  3.992 .0  4.038  0.06  .6 .0 .0
```

```
label
title-label
color
Courier 0.1 1 0       1       1        0.0 0.0 0.1
2
0 -1.00 0.0  0.0 0.0 0.8  "(C)YAMA@hakodate-ct.ac.jp"
0 -0.95 0.0  0.8 0.0 0.0  "BaTiO3 structure"


label
axis_and_coordinate_origin
color
Courier 0.05 1 0       0       1        0.0 0.0 0.1
3
10 0 0  0.8 0.8 0.8 "X-axis"
 0 10 0  0.8 0.8 0.8 "Y-axis"
 0 0 10  0.8 0.8 0.8 "Z-axis"


disjoint line
x_y_z_axis-drawing
color
6
 0 0 0  0.8 0.8 0.8
10 0 0  0.8 0.8 0.8
 0 0 0  0.8 0.8 0.8
 0 10 0  0.8 0.8 0.8
 0 0 0  0.8 0.8 0.8
 0 0 10  0.8 0.8 0.8

# x y z radius R G B


label
title-label
color
Courier 0.1 1 0       1 1        0.0 0.0 0.1
1
0 -0.85 0.0  0.8 0 0.8 "page= 1, movie data"


step2


sphere
BaTiO3_for_Ba
color
8
  .0    .0  1.0    1.35     0.2     0.2     0.6
3.992  .0  1.0    1.35     0.2     0.2     0.6
  .0  3.992 1.0    1.35     0.2     0.2     0.6
3.992 3.992 1.0    1.35     0.2     0.2     0.6

  .0    .0  5.038  1.35     0.2     0.2     0.6
3.992  .0  5.038   1.35     0.2     0.2     0.6
  .0  3.992 5.038  1.35     0.2     0.2     0.6
3.992 3.992 5.038  1.35     0.2     0.2     0.6

sphere
BaTiO3_for_O
color
6
 1.996  1.996  .0    1.40   0.9    0.9    0.9
 1.996   .0   2.019  1.40   0.9    0.9    0.9
  .0   1.996  2.019  1.40   0.9    0.9    0.9
 1.996 3.992  2.019  1.40   0.9    0.9    0.9
 3.992 1.996  2.019  1.40   0.9    0.9    0.9
 1.996 1.996  4.038  1.40   0.9    0.9    0.9
```

```
sphere
BaTiO3_for_Ti
color
1
 1.996  1.996 3.019  0.76   0.98    0     0


label
atom_position_label
color
Courier 0.05 1 0        0       1        0.0 0.0 0.1
4
  .0    .0   .0  0.8 0.8 0.8 " .0   .0   .0"
3.992  .0   .0  0.8 0.8 0.8 "3.992  .0   .0"
  .0  3.992 .0  0.8 0.8 0.8 " .0  3.992 .0"
  .0    .0  4.038 0.8 0.8 0.8 " .0   .0  4.038"


column
cylinder_2
dvertex_and_color
16
12
.0 .0 .0   .0    .0   4.038   0.06    .6 .0 .0
.0 .0 .0   .0  3.992  .0      0.06    .6 .0 .0
.0 .0 .0  3.992  .0   .0      0.06    .6 .0 .0

3.992 3.992 4.038   .0  3.992 4.038  0.06  .6 .0 .0
3.992 3.992 4.038  3.992  .0   4.038  0.06  .6 .0 .0
3.992 3.992 4.038  3.992 3.992  .0    0.06  .6 .0 .0

  .0    .0  4.038 .0 3.992 4.038  0.06  .6 .0 .0
  .0    .0  4.038 3.992 .0 4.038  0.06  .6 .0 .0
  .0  3.992  .0  .0 3.992 4.038  0.06  .6 .0 .0
  .0  3.992  .0  3.992 3.992 .0   0.06  .6 .0 .0
3.992  .0    .0  3.992 3.992 .0   0.06  .6 .0 .0
3.992  .0    .0  3.992 .0  4.038  0.06  .6 .0 .0

label
title-label
color
Courier 0.1 1 0       1       1        0.0 0.0 0.1
2
0 -1.00 0.0  0.0 0.0 0.8  "(C)YAMA@hakodate-ct.ac.jp"
0 -0.95 0.0  0.8 0.0 0.0  "BaTiO3 structure"

label
axis_and_coordinate_origin
color
Courier 0.05 1 0       0       1        0.0 0.0 0.1
3
10 0 0  0.8 0.8 0.8 "X-axis"
 0 10 0  0.8 0.8 0.8 "Y-axis"
 0 0 10  0.8 0.8 0.8 "Z-axis"

disjoint line
x_y_z_axis-drawing
color
6
 0 0 0  0.8 0.8 0.8
10 0 0  0.8 0.8 0.8
 0 0 0  0.8 0.8 0.8
 0 10 0  0.8 0.8 0.8
 0 0 0  0.8 0.8 0.8
 0 0 10  0.8 0.8 0.8

# x y z radius R G B


label
title-label
color
Courier 0.1 1 0       1 1        0.0 0.0 0.1
1
0 -0.85 0.0  0.8 0 0.8 "page= 2, movie data"
```

Fig.10. Schematic data listed representation as summarized appendix of a almost final purchase product of the 4D "movie.mgf"that through a purchase procedure via awk script of Fig 3 and with the seed of static 3D MGFs of Fig 2..

SMP-Based Approach for Intelligent Service Interaction

Hosam AlHakami
(STRL)
De Montfort University
Leicester, United Kingdom
Email: hosam.alhakami@myemail.dmu.ac.uk

Feng Chen
(STRL)
De Montfort University
Leicester, United Kingdom
Email: fengchen@dmu.ac.uk

Helge Janicke
(STRL)
De Montfort University
Leicester, United Kingdom
Email: heljanic@dmu.ac.uk

Abstract—**Service-oriented computing is establishing itself as a new computing paradigm in which services advertise their capabilities within a network, and then are used, composed and orchestrated by other services and end-users. Whilst many approaches to matching service providers with consumers of their services have been developed in the past, the proposed approach in this paper takes a different view of the problem in that it does not look for the fittest individual utility, but views it as a constrained optimisation problem that matches between a set of services and a set of service requests. Our approach addresses this problem using an adaptation of the well known stable-marriage problem and demonstrates how matching between services and requests to a certain threshold can be expressed. This will contribute a fair assignment between services and requests based on their preferences. As the current state of the service selection process considers only the view of requests, the proposed approach can ensure several features to the services such as service protection and service quality, e.g. it can ensure the preservation of service availability by redirecting a coming request to a similar service if the current service is busy.**

Keywords—*Service Selection; Stable Marriage Problem (SMP); Selective Strategy*

I. Introduction

Recent research of service selection focuses on matching behaviour, trying to interact with services in a smart way. This can positively impact several relevant big topics on service-oriented computing, such as cloud computing which provides both software and resources as services, and service composition in which aggregation of services are composed to automate a certain job.

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

In this paper we will apply a variant of the stable marriage problem to service oriented computing (SOC) and use it to match between service providers and service consumers. Current selection algorithms take the QoS (Quality of Services) constraints of the requestor (service consumer) into account, but do not consider the preferences of the service provider in the process. Our work expands this to include preferences that a service may have towards service requests. Example preferences of a service may include returning customers, payment options and country of origin. The web service selection process should therefore consider the values of both parties.

The remainder of this paper is structured as follows. In Section II we provide the context of Service Selection. In Section III we introduce the background to the SMP research. In Section IV we present an adapted SMP algorithm that is suitable to generate fair and stable matches between services and requests. In Section V we deeply discuss our approach and its expected performance with a hotel booking case study. In Section VI we conclude the paper and outline future work in this area.

II. Service Selection

Service selection is the process of choosing the appropriate service which fulfils the client's requirement. Service selection is a key step to affect the whole process of service composition. Different approaches have been applied to match between service providers and consumers. Typically the selection relies on the search function of Universal Description, Discovery and Integration (UDDI) which often provides a relatively poor search facility by allowing only a keyword based search of web services. Consequently, returning unrelated services since UDDI registry encompasses both checked and unchecked services [2]. Huang et al. in [3] present an efficient service selection scheme that considers non-functional characteristics while choosing web services by service requesters. Their model considers different data types, and uses multiple criteria decision making (MCDM) with weighted sum model (WSM) to help service requesters evaluate services numerically. They proof using experimental results that their service selection scheme performs much better than the enumerative method, and outperforms other related methods in terms of efficiency and effectiveness. The Figure 1 below illustrates the whole process through the three major stages of Huang's model.

Fig. 1. Service Selection process [3].

Various definitions of Quality of Service (QoS) has been

provided previously; Schmidt in [4] defines the QoS as The degree to which a system, component or process meets customer or user needs or expectations. Alrifai and Risse [5] propose a solution that combines global optimisation with local selection techniques to prevent local selection strategy fails in handling global QoS requirements. They use distributed local selection to find the best web services that satisfy these local constraints. Baldoni et al. [6] show that the semantic matchmaking techniques used in the services is lacking preservation and they provide an approach to overcome these limitations. Yu et al. [7] present a broker-based architecture to facilitate the selection of QoS-based services. Recent approaches of service selection are deeply involved in semantic manner using ontology which draw the attention of many researchers [8] [9] [10] [11] [12].

Discovering the recommended web services needed by users among the vast resources is a significant challenge. Zhong et al. [13] discuss functional attributes and non-functional attributes of the web services resources from the resources balanced perception of FQoS (Functional Quality of Service) and QoS. They propose a web service description, discovery and selection mechanism towards balanced perceptions of FQoS and QoS. In this mechanism, FQoS is more used to describe and discover services, and QoS is more used to select services. Due to the perspective differences, they point out several problems on the research on service functions and quality of service. There are big differences among QoS indicators. Due to the similarity between the functional and non-functional attributes, there is an ambiguity of specifying service functions out of QoS [14]. Figure 2 below illustrates a brief depiction about FQoS.
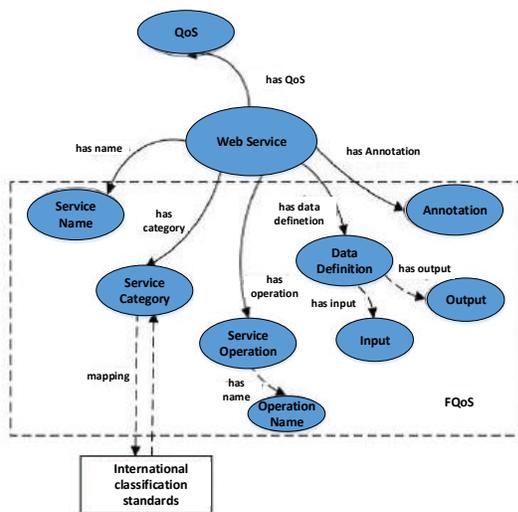

Fig. 2. Description Framework Of FQoS [13].

The main purpose of designing the Clouds is to provide services to users and providers needs to be made up for sharing their capability and resources [15]. Armbrust et al. [16] discuss many obstacles of growth cloud computing paired with the suggested solutions (opportunities) to overcome these issues as depicted in Table I. Also, many organisation worry about the availability of the computing services, which therefore affect the cloud computing workflow. Some software as a service (SaaS) products set a high standard in this regard to prevent

the issue of service unavailability such as Google Search [16]. Table I shows the services outages and the causes, for Amazon Simple Storage Service (S3), AppEngine and Gmail in 2008.

TABLE I. OBSTACLES TO AND OPPORTUNITIES FOR GROWTH OF CLOUD COMPUTING. [13]

Service and Outage	Duration	Date
S3 outage: authentication service overload leading to unavailability	2 hours	2/15/08
S3 outage: Single bit error leading to gossip protocol blowup	6-8 hours	7/20/08
AppEngine partial outage: programming error	5 hours	6/17/08
Gmail: site unavailable due to outage in contacts systems	1.5 hours	8/11/08

As services providers need to meet the requirement of the consumers, providing a mechanism of matching the services description to the required services is very significant. Meanwhile, massive services of the Internet need to assign the services to the requesters efficiently and automatically. Therefore, providing service matching approach can tackle above issue by redirecting the related request to the most similar available services when the current service is busy. In this paper SMP algorithm will be investigated to improve the process of service selection.

III. SMP ALGORITHM

A. Classic SMP Algorithm

In 1962, David Gale and Lloyd Shapley published their paper College admissions and the stability of marriage [1]. This paper was the first to formally define the Stable Marriage Problem (SMP), and provide an algorithm for its solution. The SMP is a mechanism that is used to match two sets of the same size, considering preference lists in which each element expresses its preference over the participants of the element in the opposite set [1]. Thus, the output has to be stable, which means that the matched pair is satisfied and both candidates have no incentive to disconnect.

A matching M in the original SMP algorithm is a one-to-one correspondence between the men and women. If man m and woman w are matched in M, then m and w are called partner in M, and written as $m = PM(w)$ (which is the M-partner of w), $w = PM(m)$ (the M-partner of m). A man m and a woman w are said to block a matching M, or called a blocking pairs for M if m and w are not partners in M, but m prefers w to $PM(m)$ and w prefers m to $m = PM(w)$ [17]. Therefore, a matching M is stable when all participants have acceptable partners and there is no possibility of forming blocking pairs. This problem is in interest of a lot of researchers in many different areas from several aspects.

An instance I of SM involves n men and m women, each of whom ranks all n members of the opposite sex in strict order of preference. In I we denote the set of men by $m = m1, m2, mn$ and the set of women by $w = w1, w2, ...wn$. In SM the preference lists are said to be complete, that is each member of I ranks every member of the opposite sex as depicted in Figure 3.

There are some derived versions of the original SMP algorithm are considered such as SMI (Stable Marriage with incomplete list) to manage problems of unacceptable partner and different set sizes.
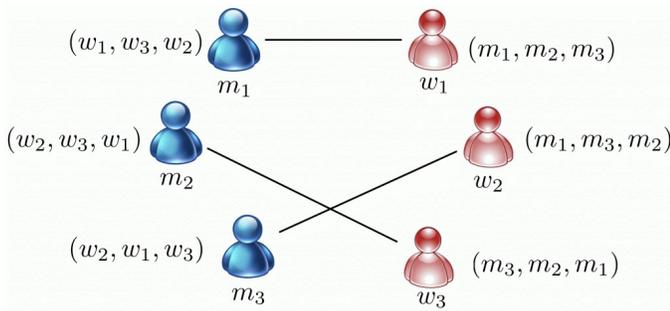
Fig. 3.   General view of SMP. [18]

Fig. 4.   A stable marriage instance of size 4.

B. Gale Shapley Extended Algorithm

The algorithm presented by Gale and Shapley for finding a stable matching uses a simple deferred acceptance strategy, comprising proposals and rejections. There are two possible orientations, depending on who makes the proposals, namely the man-oriented algorithm and the woman-oriented algorithm. In the man-oriented algorithm, each man m proposes in turn to the first woman w on his list to whom he has not previously proposed. If w is free, then she becomes engaged to m. Otherwise, if w prefers m to her current fiancé m, she rejects m, who becomes free, and w becomes engaged to m. Otherwise w prefers her current fiancé to m, in which case w rejects m, and m remains free. This process is repeated while some man remains free. For the woman-oriented algorithm the process is similar, only here the proposals are made by the women.

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

Theorem 1 All possible execution of the Gale-Shapley algorithm (with the men as proposers) yields the same stable matching, and in this stable matching, each man has the best partner that he can have in any stable matching [17].

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

Theorem 2 In the man-optimal stable matching, each woman has the worst partner that she can have in any stable matching [17].

The following example in Figure 4 gives different output for both man-optimal and woman optimal for the same instance which formed out of 4 elements.
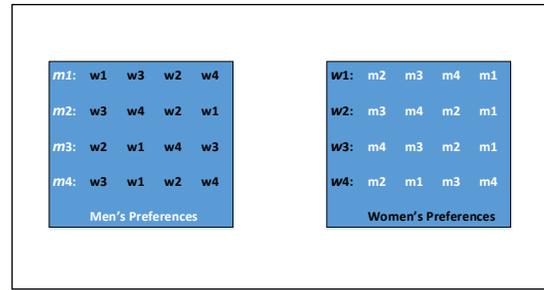
The stable matching generated by both man-oriented and women-oriented versions is: $Man - oriented = M0 = (1,4),(2,3),(3,2),(4,1)$. $Woman - oriented = Mz = (1,4),(2,1),(3,2),(4,4)$. An extended version of Gale-Shapley algorithm has been designed to improve the basic algorithm. The extended version reduces the preference list by eliminating specific pairs that can be clearly identified as unrelated to any stable matching. The deletion process of such pair is performed by deleting each other from the preference lists.

Algorithm 1 Extended Gale-Shapley algorithm [17]

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

The main SMP algorithm performs a one-to-one cardinality over the matched candidates. Furthermore, the relationship of one-to-many is presented in one of the extended SMP algorithms which called Hospital resident (HR), basically many residents assigned to one hospital [19]. Whereas in this approach a many-to-many relationship is needed. Hence, we extended the original SMP algorithm to cope with the new requirements of this proposed approach. We will build on this background in section IV where we use the extended Gale-Shapley algorithm in our application to the problem of service selection.

Applying SMP algorithm in services needs first building the preference lists of both services. This can be achieved using service profile to specify the most similar related participants (services in the other pool). Thus, each service from $p1$ needs to strictly order the services from $p2$ based on the similarity and vice versa. However, the original SMP algorithm performs a single assignment (one-to-one) for the involved candidates, which at some points does not help especially in the case of allocating more than one service to the right service for example. Figure 5 below illustrates a brief depiction about matching based service selection.
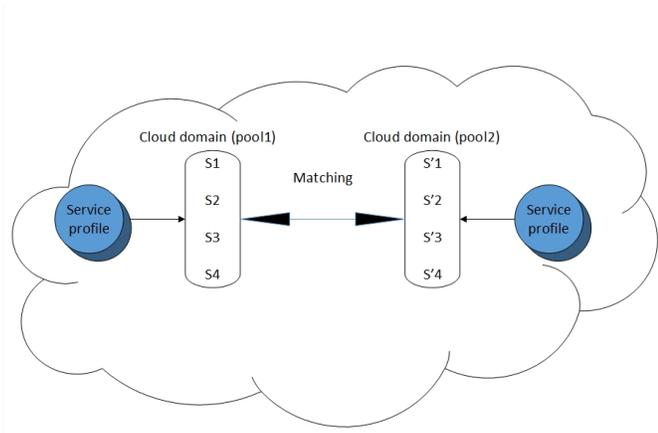
Fig. 5.   Service Matching in Cloud [13].

IV.   EXTENDED SMP ALGORITHM (DUAL-PROPOSED) FOR SERVICE SELECTION

A. Dual-Proposed Scheme

SMP has solved several similar optimisation issues in different fields such as matching jobs to the most suitable job-seekers. Since the SMP algorithm allows only the candidates of the first set (Men) to propose to their first choices, this research devotes to increase the fairness of SMP by allowing the candidates of the second set (Women) to make their own choices i.e. proposes to the best of their choices of the opposite set . To allow the candidates from both sets, dual-proposed scheme should be applied to propose for their preferences. This slight amendment has enhanced the precision of the matching process; it is illustrated in Figure 6 below. In the main SMP algorithm the desire is not controlled by the similarity, thus the assigned candidates are not meant that they are similar to each other. However, in service selection the concept of similarity plays an important role. Therefore, aforementioned extension of the current state of SMP is necessary to be effectively applied in such applications.

A novel matching scheme is needed to achieve smart interaction between the services and their corresponding requests. This led to a quality satisfaction for selecting the appropriate services to the equivalent requests. This improves the current relation in the original SMP algorithm which bases its matching process on the desire of the candidates of the first set regardless the similarity factor. Here the similarity has the same meaning as desire in the original SMP algorithm.
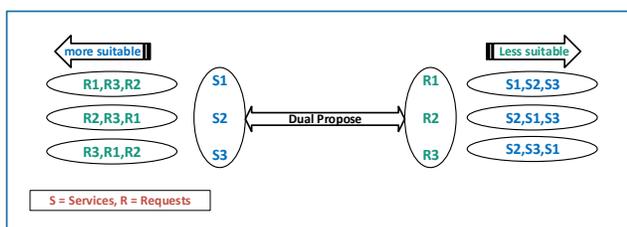

Fig. 6.   Dual-Proposed Algorithm.

Practically, this process gives two different stable matched

pairs; respectively man-optimal and woman-optimal. Thus, we enclose a novel way of assigning men and women to each other by adding a selective strategy. This strategy helps to choose the most optimal pair among the final result. For service selection dual-proposed is matching services from different domains that take into account the preferences of both parties. This means that our approach is able to develop match making services that trying to optimise the pairings from both perspectives fairly.

This presented matching scheme covers a noticeable hole of the current state of the original SMP algorithm and perfectly assigned to the world of software engineering, serving in several aspects the needed allocation process that requires a high consideration of both sides of matched candidates, for example the approach presented by AlHakami et al. in [20] to enhance the process of detecting cloned software. Also, our approach shows the ability to enhance the quality of allocating the similar services to each other, through considering the desire (similar or equivalence) of the matched services, which result in increased the satisfaction of the candidates in each pair, expecting a high stability. However, the Scheme should expect some limitation in terms of the time feature, as the consideration of the candidates' desire needs more computation and recursion to fulfil and reached the highly required stability.

B. Dual-Proposed Algorithm

The dual-proposed algorithm acts as if it were both man-optimal and woman-optimal. The combination of these two versions that supplemented with a selective strategy using love's degree factor, results in complete concrete dual-proposed scheme. The algorithm 2 of dual-proposed scheme consists of two phases followed by the Selective Strategy as following:

- **Phase 1 man-optimal algorithm.**

- **Phase 2 woman-optimal algorithm.**

- **Apply selective strategy.**

C. Selective Strategy

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

The selective strategy or choosy strategy as presented in the previously [21], determines two main factors, respectively, love's degree and contrast's degree. Love's degree reflects the degree of love from the view of both involved candidates (services etc.). To converge these views, we add the degree of love for both of participants (in the same pair) candidates and divide the result by two. The final result is the love's degree of the pair. The contrast's degree reflects the difference between the actual love's degree of the involved candidates (services etc.). Thus, the most preferable pair is that with small difference in its contract's degree. This factor helps when two different pairs have the same love's degree. Also, when more than one candidate has the same love's degree with a certain candidate, then the right candidate will be chosen. Figure 7

Algorithm 2 Dual-Proposed Algorithm

1: assign each person to be free
2: **while** some man m is free **do**
3: begin
4: w:= first woman on m's list;
5: **if** some man p is engaged to w **then**
6: assign p to be free;
7: assign m and w to be engaged to each other;
8: **for each** successor m' of m on w' list **do**
9: delete the pair(m',w)
10: end;
11: assign $M1$ to the man-optimal pair;
12: assign each person to be free
13: **while** some woman w is free **do**
14: begin
15: m:= first man on w's list;
16: **if** some woman s is engaged to m **then**
17: assign s to be free;
18: assign w and m to be engaged to each other;
19: **for each** successor w' of w on m' list **do**
20: delete the pair(w',m)
21: end;
22: assign $M2$ to the woman-optimal pair;
23: apply selective strategy on $M1$ and $M2$
24: end;

depicts the calculation of selective strategy upon both services and requests respectively numbers and letters.
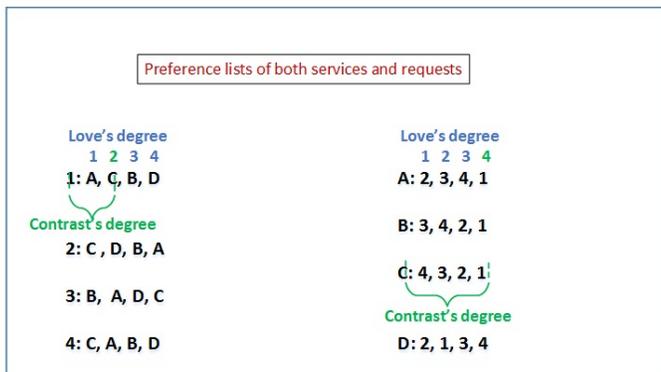

Fig. 7. Selective Strategy Scheme

With regards to the previous example in Figure 4 the new results after apply the selective strategy are as following:

TABLE II.   NEW MAN-ORIENTED

M0	Love's degree	Contrast's degree	Weight
(1,D)	2	0	1
(2,C)	3	3	3
(3,B)	2	1	2
(4,A)	2	1	2

$ServiceView = New - Man - oriented = (1,D),(2,A),(3,B),(4,A)$.

TABLE III.   NEW WOMAN-ORIENTED

Mz	Love's degree	Contrast's degree	Weight
(1,D)	2	0	1
(2,A)	2	1	2
(3,B)	2	1	2
(4,C)	3	1	2

$RequestView = New - Woman - oriented = (1,D),(2,A),(3,B),(4,A),(4,C)$. The combination of both new versions, each candidate may be assigned to more than one candidates and this based upon the features considered by the selective strategy.

Similarly The following example in Figure 8 considers the selective strategy with different symbols. This example shows different output from the main SMP algorithm, which allows a candidate to be assigned to more than one candidate from the opposite set of participants, which reflects the cardinality of many-to-many. The new algorithm provides different matching pairs from the original man-optimal algorithm.
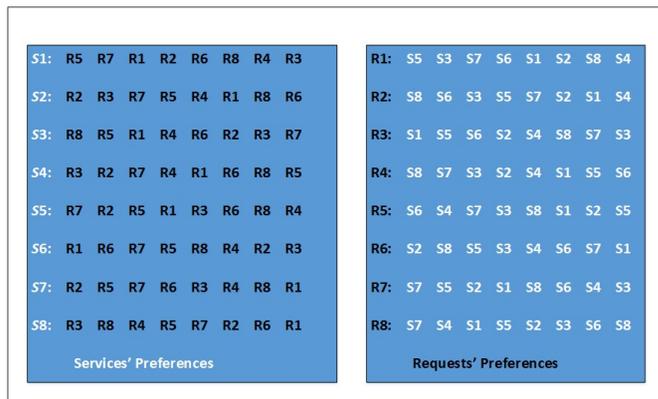
Fig. 8. Service marriages (Similarity-based).

M0 = { $(S1,R5)$, $(S2,R3)$, $(S3,R8)$, $(S4,R6)$, $(S5,R7)$, $(S6,R1)$, $(S7,R2)$, $(S8,R4)$}

TABLE IV.   SERVICE-VIEW

Pairs (m,w)	Love's degree	Contrast's degree	Weight
$(S1,R5)$	1	5	3
$(S2,R3)$	2	2	2
$(S3,R8)$	1	5	3
$(S4,R6)$	6	1	4
$(S5,R7)$	1	1	1
$(S6,R1)$	1	3	2
$(S7,R2)$	1	4	3
$(S8,R4)$	3	2	3

So, the output after applying our strategy is as following: New-Service-Optimal= { $(S1,R5)$, $(S2,R3)$, $(S3,R2)$, $(S4,R8)$, $(S5,R7)$, $(S6,R1)$, $(S7,R7)$, $(S8,R4)$ }.

Mz= { $(S1,R3)$, $(S2,R6)$, $(S3,R2)$, $(S4,R8)$, $(S5,R1)$, $(S6,R5)$, $(S7,R7)$, $(S8,R4)$}

Pairs (m,w)	Love's degree	Contrast's degree	Weight
$(R1,S5)$	1	3	2
$(R2,S3)$	3	3	3
$(R3,S1)$	1	7	4
$(R4,S8)$	1	2	2
$(R5,S6)$	1	3	2
$(R6,S2)$	1	7	4
$(R7,S7)$	1	2	2
$(R8,S4)$	2	5	4

So, the output after applying our strategy is as following: New-Request-Optimal= { $(R1,S5)$, $(R1,S6)$, $(R2,S3)$, $(R3,S2)$, $(R4,S8)$, $(R5,S6)$, $(R6,S4)$, $(R7,S5)$, $(R8,S3)$ }. = { $(S2,R3)$, $(S3,R2)$, $(S3,R8)$, $(S4,R6)$, $(S5,R1)$, $(S5,R7)$, $(S6,R1)$, $(S6,R5)$, $(S8,R4)$ }.

From above results we can derive the dual-propose-multi-allowance as following:

Service-view $\bigcup$ Request-view = { $(S1,R5)$, $(S2,R3)$, $(S3,R2)$, $(S3,R8)$, $(S4,R6)$, $(S4,R8)$, $(S5,R1)$, $(S5,R7)$, $(S6,R1)$, $(S6,R5)$, $(S7,R7)$, $(S8,R4)$ }.

Whereas, Service-view $\bigcap$ Request-view = { $(S2,R3)$, $(S3,R2)$, $(S5,R7)$, $(S6,R1)$ }.

V. CASE STUDY

To apply the SMP algorithm in service selection, it needs first to build the preference lists of both requests and services. This can be achieved using service profile to specify the most similar related participants (services or requests). Each request needs to strictly order the service based on the similarity and vice versa (similar to Figure 8). The traditional SMP algorithm performs a single assignment (one-to-one) for the involved candidates, which does not help especially in the case of allocating more than one service requests to the right service for example. Dual-proposed algorithm has been proposed to fulfil this requirement.

A. A Hotel Reservation Example

Figure 9 is a general example of web service composition scenario.
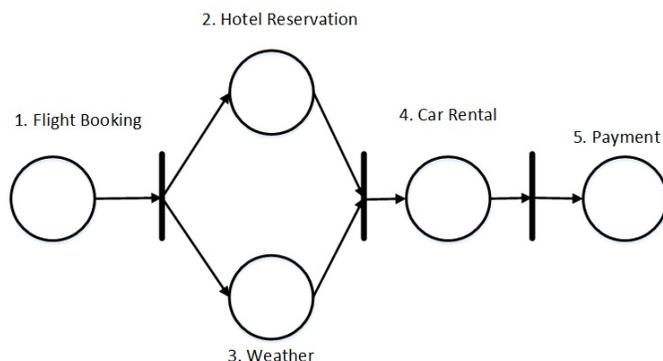
Fig. 9. General Example of web services.

Figure 10 shows an engaged service, hotel reservation, to the most suitable requests using SMP algorithm based on the QoS and the values of both candidates. There are different hotel services to show the rooms available, the style of the rooms etc. There are also some hotel reservation requests from the customers or agents. Then the extended dual-proposed algorithm can be applied to choose the best services.
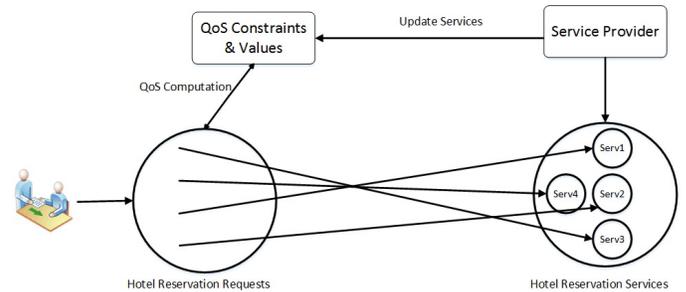

Fig. 10. SMP-based service selection.

The result is matching of service providers (hotels) and consumers (guests) that do not only take into account the preferences of the respective consumers, but also those of the providers. This means that we are now able to develop match making services that not only decide on the basis of the request's preferences, but are actually trying to, within the current set of services requests and services, optimise the pairings from both perspectives fairly.

B. Discussion

The network administrator should act as the stakeholder and take the advantage of monitoring the process of matching services to requests. This provides reliable information such as profile and history details of both services and requests. The result is a service interaction in which optimal assignments of these two parties ensure safety, reliability, dependability and stability to their candidates.

Based on the characteristics of both services and requirements of customers' requests, our approach can start triggering the matching process to allocate the most proper customers' requests to the most appropriate services and considers both views using our smart selective strategy.

Our approach contributes to the service selection by adding the matching behaviour rather than the traditional existing search-based techniques which picking up the required services from the UDDI directory regardless the constraints of the chosen services.

Web services should be dealt with as independent objects that have their own view about incoming consumers' request and are aware of its requirements. Therefore, SMP-based approach gives services rights to whether accept or decline the customers' requests, with regards to their conditions. This can corroborate the concept of equal opportunity which in turn ensures the stability.

Moreover, this approach helps services to check the trust level of the client, which supports both safety and reliability. On demand, the service may choose to reject the client because he demanding highly computational task that may impact other ongoing tasks in the service to meet a quality of service that it promised to other clients.

The following Diagrams 11 and 12 show some considered facets of both consumer's request and web service respectively, in a simple prototype example of hotel reservation.
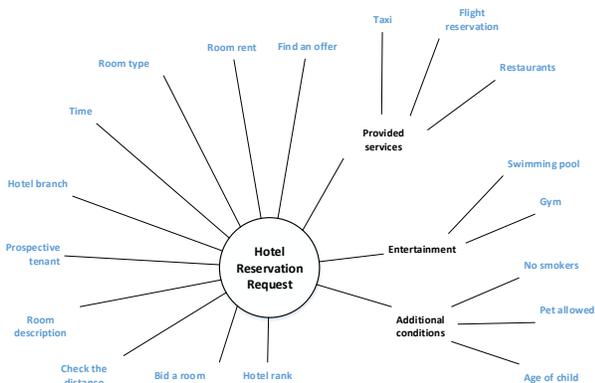


Fig. 11. Hotel Reservation Request.
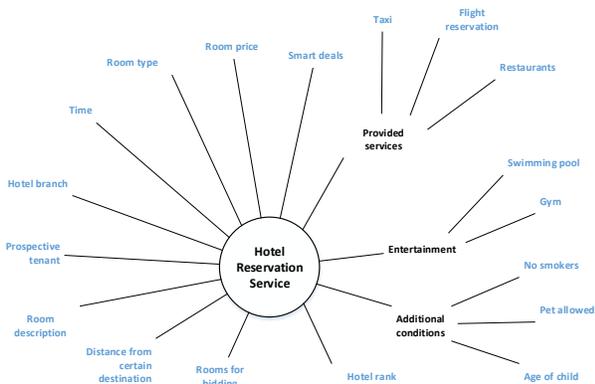
Fig. 12. Hotel Reservation Service.

The following Figure 13 depicts an early pre-step of matching consumer's requests to the services of hotel reservation using SMP-Based approach.

VI. CONCLUSION

Stable marriage problem is well-known common matching algorithms, helps in several applications in different aspects of live for instance assigning medical schools graduates students to the most suitable hospitals. The paper presents a newly crucial extension which effectively touches a wide range of software engineering fields such as service selection. Our contribution in this paper is the selective strategy, that compromises between the preferences of the service provider and the service request in a service oriented computing environment. Our scheme can increase the quality of allocating the services to the corresponding requests, through considering the desire of the matched candidates, which results in increased satisfaction of the candidates in each pair. This leads to more stable service bindings and has the potential to reduce inefficiencies that are incurred through re-binding and alternative lookups. However, the proposed scheme has some limitations in terms
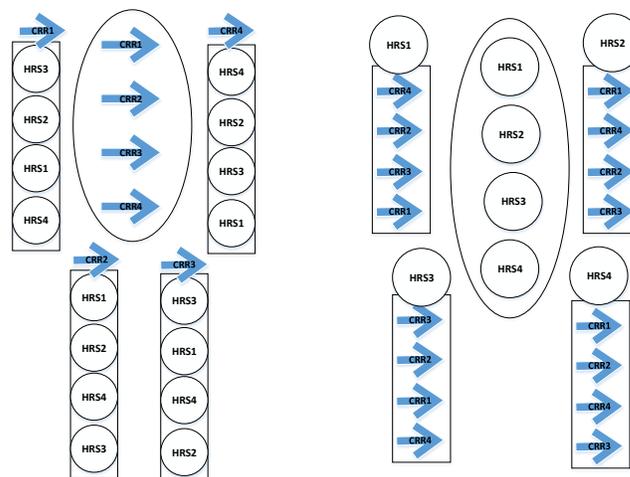

Fig. 13. Hotel Reservation Example.

of its complexity and would require longer time to reach the highly required stability. In our future work we would like to look at how dynamic constraints such as incurred by load and execution constraints on the providers would influence the stability of the matches. We also plan to evaluate our approach against strategies that consider only the interests of the service consumer.

REFERENCES

[1] D. Gale and L. Shapley, "College admissions and the stability of marriage," *American Mathematical Monthly*, pp. 9–15, 1962.

[2] M. Jiang, Z. Ding, and J. Liu, "Service selection based on behavior matching," *Journal of Software*, vol. 7, no. 9, pp. 1950–1959, 2012.

[3] A. F. Huang, C.-W. Lan, and S. J. Yang, "An optimal qos-based web service selection scheme," *Information Sciences*, vol. 179, no. 19, pp. 3309–3322, 2009.

[4] M. E. Schmidt, *Implementing the IEEE software engineering standards*. Sams, 2000.

[5] M. Alrifai and T. Risse, "Combining global optimization with local selection for efficient qos-aware service composition," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 881–890.

[6] M. Baldoni, C. Baroglio, A. Martelli, V. Patti, and C. Schifanella, "Service selection by choreography-driven matching," in *Emerging Web Services Technology, Volume II*. Springer, 2008, pp. 5–22.

[7] T. Yu, Y. Zhang, and K.-J. Lin, "Efficient algorithms for web services selection with end-to-end qos constraints," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, p. 6, 2007.

[8] P. A. Khutade and R. Phalnikar, "Qos aware web service selection and ranking framework based on ontology," *International Journal of Soft Computing and Engineering*, vol. 4, no. 3, pp. 77–81, 2014.

[9] W. Nagy, H. M. Mokhtar, and A. El-Bastawissy, "A flexible tool for web service selection in service oriented architecture," *International Journal of Advanced Computer Science and Applications*, vol. 2, no. 12, pp. 191–201, 2011.

[10] G. Guo, F. Yu, Z. Chen, and D. Xie, "A method for semantic web service selection based on qos ontology," *Journal of Computers*, vol. 6, no. 2, pp. 377–386, 2011.

[11] W. Junhao, G. Jianan, J. Zhuo, and Z. Yijiao, "Semantic web service selection algorithm based on qos ontology," in *Service Sciences (IJCSS), 2011 International Joint Conference on*. IEEE, 2011, pp. 163–167.

[12] N. Keskes, A. Lehireche, and A. Rahmoun, "Web services selection based on context ontology and quality of services." *Int. Arab J. e-Technol.*, vol. 1, no. 3, pp. 98–105, 2010.

[13] W. Zhong, N. Guihua, C. Donglin, and Z. Peilu, "Web service description framework and selection mechanism towards balanced perceptions of fqos and qos," *Theoretical and Applied Information Technology*, vol. Vol. 47 No.1, 2013.

[14] C.-H. Tao and Z.-Y. Feng, "Novel qos-aware web service recommendation model," *Application Research of Computers*, vol. 10, p. 083, 2010.

[15] H. AlHakami, H. Aldabbas, and T. Alwadan, "Comparison between cloud and grid computing: review paper," *International Journal on Cloud Computing: Services and Architecture*, vol. 2, no. 4, pp. 1–21, 2012.

[16] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica *et al.*, "A view of cloud computing," *Communications of the ACM*, vol. 53, no. 4, pp. 50–58, 2010.

[17] D. Gusfield and R. W. Irving, *The stable marriage problem: structure and algorithms*, 1989.

[18] *http://iqua.ece.toronto.edu/spotlights/matching/*, Std. [Online]. Available: http://iqua.ece.toronto.edu/spotlights/matching/

[19] C. Cheng, E. McDermid, and I. Suzuki, "A unified approach to finding good stable matchings in the hospitals/residents setting," *Theoretical Computer Science*, vol. 400, no. 1, pp. 84–99, 2008.

[20] H. AlHakami, F. Chen, and H. Janicke, "An extended stable marriage problem algorithm for clone detection," *International Journal of Software Engineering & Applications*, vol. 5, no. 4, pp. 103–122, 2014.

[21] ——, "Smp-based service matching," in *Science and Information Conference (SAI), 2014*. IEEE, 2014, pp. 620–625.

Reducing the Correlation Processing Time by Using a Novel Intrusion Alert Correlation Model

Huwaida Tagelsir Ibrahim Elshoush
Computer Science Department
Faculty of Mathematical Sciences
University of Khartoum - Sudan
Email: htelshoush@uofk.edu

Abstract—Alert correlation analyzes the alerts from one or more Collaborative Intrusion Detection Systems (CIDSs) to produce a concise overview of security-related activity on the network. The correlation process consists of multiple components, each responsible for a different aspect of the overall correlation goal. The sequential order of the correlation components affects the correlation process performance. Furthermore, the total time needed for the whole process depends on the number of processed alerts in each component. This paper presents an innovative alert correlation framework that minimizes the number of processed alerts on each component and thus reducing the correlation processing time. By reordering the components, the introduced correlation model reduces the number of processed alerts as early as possible by discarding the irrelevant, unreal and false alerts in the early phases of the correlation process. A new component, *shushing the alerts*, is added to deal with the unrelated and false positive alerts. A modified algorithm for fusing the alerts is outlined. The intruders' intention is grouped into attack scenarios and thus used to detect future attacks. DARPA 2000 intrusion detection scenario specific datasets and a testbed network were used to evaluate the innovative alert correlation model. Comparisons with a previous correlation system were performed. The results of processing these datasets and recognizing the attack patterns demonstrated the potential of the improved correlation model and gave favorable results.

Keywords — Alert Correlation, Alert Reduction, Intrusion Detection Systems, False Alarm Rate

I. INTRODUCTION

Intrusion Detection Systems (IDSs) play an essential role in minimizing the damage caused by different attacks. On the other hand, many of the weaknesses in traditional IDSs are due to the lack of collaborations among different detection mechanisms, and between intrusion detection and other network management operations and security mechanisms. Therefore, a Collaborative Intrusion Detection System (CIDS) architecture is introduced. In particular, a Collaborative Intelligent Intrusion Detection System (CIIDS) is proposed to include both misuse- and anomaly-based techniques, since it is concluded from recent research that the performance of an individual

detection engine is rarely satisfactory. Employing multiple IDSs and other security systems gives a better view of the monitored network. They may cooperate to complement each other's coverage. Even when different detection methods are used, they analyze each other's alerts and reduce false positive alerts [1][2]. It has been proven by many researchers that collaborative approaches are more powerful and give better performance over individual approaches. In fact, the use of complementary IDSs is a promising technique as each IDS implements different detection scheme, algorithms and signatures and therefore gives a more exact and complete view of suspicious network events. Specifically, two main challenges in current CIDSs research are highlighted: CIDS architectures and alert correlation algorithms. *Alert correlation* in CIDSs will be more challenging. Deploying multiple IDSs might generate a huge number of alerts, where many are redundant, irrelevant and false positive alerts. Hence, data reduction, such as alert aggregation, alert filtering and false alert reduction, without losing valuable information is essential. The focus in this research is on correlation of Collaborative Intelligent Intrusion Detection System (CIIDS) alerts.

Automation of alert management and analysis is crucial because of the huge number of alarms, the false positives and the irrelevant alarms and to study the cause of these alarms. Thus, there is a need of alert correlation, which is a process that contains multiple components with the purpose of analyzing alerts and providing high-level insight view on the security state of the network surveillance [1][3][4]. Correlation aims to relate a group of alerts to build a big picture of the attacks, hence can be used to trace an attack to its source.

The core of this process consists of components that implement specific function, which operate on different spatial and temporal properties [5].

The correlation components are effective in achieving alert reduction and abstraction. Research shows that the effectiveness of each component depends heavily on the nature of the data set analyzed [5]. Moreover, the performance of the correlation process is significantly influenced by the topology of the network, the characteristics of the attack, and the available meta-data [6].

Since alerts can refer to different kinds of attacks at different levels of granularity, the correlation process cannot treat all
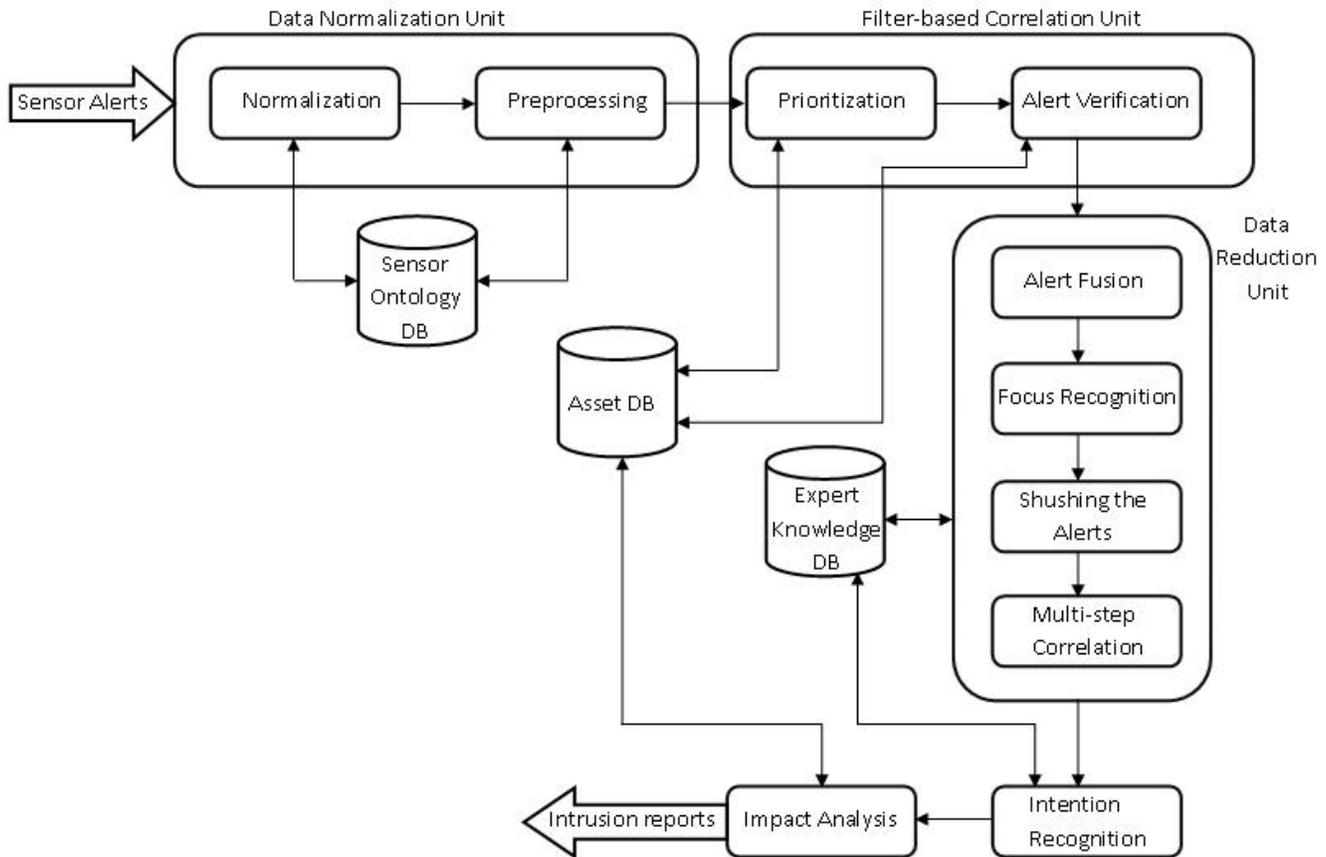
Fig. 1: The Innovative Correlation Model [7].

alerts equally. Instead, it is necessary to have a set of components that focus on different aspects of the overall correlation task. Some components, see Fig.1, e.g. those at the initial and second units, implement general functionality applicable to all alerts, independent of their type. Other components (e.g. in the third unit) are responsible for performing specific correlation tasks that cannot be generalized for arbitrary alerts, but for certain class of alerts.

Thus, one cannot, in general, determine a ranking among components with respect to their effectiveness. Each component can contribute to the overall analysis. Therefore, the most complete set of components should be used [5].

An innovative framework focuses on reordering the correlation components such that redundant, irrelevant and false alerts are reduced as early as possible thus reducing the number of processed alerts to enhance the performance. The unrelated alerts that are not correlated are dealt with in a separate component, *shushing the alerts*. Hence, the overall effectiveness of the correlation process is improved.

II. Overview of Improved Alert Correlation Framework

The proposed architecture, see Fig. 1, is composed of ten components: *normalization*, *preprocessing*, *prioritization*, *alert verification*, *alert fusion*, *focus recognition*, *shushing the*

alerts, *multi-step correlation*, *intention recognition*, and *impact analysis* [7].

In the *normalization* component, alerts that are generated by multiple IDSs are collected and stored in a database before they are modeled and converted into a standard format called Intrusion Detection Message Exchange Format (IDMEF) [8]. Then data *preprocessing* is required in order to clean the data, do feature extraction and selection, and finally deal with any incomplete or missing data [9][10][11][12].

The *filter-based correlation* unit either assigns a priority to each alert or identifies irrelevant alerts. Thus, alerts are ranked based on their severity level in order to distinguish between the high and low risks alerts depending on information in the asset DB. In the *alert verification* component, alerts are checked to find out the verifiable alerts, false positives and unverifiable alerts.

Redundant alerts are fused based on similarity functions [2] in the *alert fusion* component in the *data reduction unit*. This component combines a series of alerts that refer to attacks launched by one attacker against a single target. It removes duplicates created by the independent detection of the same attack by different sensors, and also correlates alerts that are caused by an attacker who tests different exploits against a certain program or that runs the same exploit multiple times to guess correct values for certain parameters (e.g., the offsets

and memory addresses for a buffer overflow) [5][6][17].

The alert fusion component method, see algorithm 1 below, keeps a sliding timewindow of alerts. The alerts within the timewindow are stored in a time-ordered queue. When a new alert arrives, it is compared to the alerts in the queue, starting with the alert with the earliest timestamp.

A *fusion* match is found if all overlapping attributes are equal and the new alert is produced by a different sensor. Here variable *fuse* in the algorithm is set. The timestamp of the meta-alert is assigned the earlier of the sub-alerts times. The later timestamp simply indicates a delayed detection by the other sensor.

On the other hand, attack *threads* are constructed by merging alerts with equivalent source and target attributes that occur in a certain temporal proximity but the alerts need not be produced by different sensors. Hence, variable *thread* is set, see algorithm *Alert Fusion* below. The timestamp of the meta-alert is assigned the earlier of the two start-times and the later of the two end-times.

Function *fuse-merge(alert1, alert2)* in algorithm 1 assigns the different attributes to the meta-alert constructed depending on whether fuse or thread resulted.

The value of the time window should be a good trade-off between a small value, which would cause several attack threads to go undetected, and a larger value, which would slow down the system by requiring the component to keep a large number of alerts in the queue.

Following is algorithm 1, the alert fusion component method.

Algorithm 1:
Alert Fusion
Parameter *window-size, fuse-window, thread-window*
Global *alert-queue, fuse, thread*

fuse ← *false*
thread ← *false*

fuse(alert)
al ← **get** *a:alert with lowest start-time* **from** *alert-queue* **where**
if *alert.analyzer* ∩ *a.analyzer is empty* **and** *all overlapping attributes except start-time, end-time, analyzer, alertid are equal* **then**
    fuse
    window-size = *fuse-window*
else
    if *alert.victimhosts* = *a.victimhosts* **and** *alert.attackerhosts* = *a.attackerhosts* **then**
        thread
        window-size = *thread-window*
    end if
end if
if *al¬null* **then**
    replace *al in alert-queue with fuse-merge(alert, al)*
else
    add *alert to alert-queue*
    remove *all a:alert* **from** *alert-queue* **where**
    a.start-time < *(alert.start-time - window-size)*
    pass *removed alerts to next correlation component*
end if

fuse-merge(alert1, alert2)
r ← **new** *alert*
r.alertid ← *get unique-id()*
r.start-time ← *min(alert1.start-time, alert2.start-time)*
r.reference ← *(alert1.alertid* ∪ *alert2.alertid)*
if *fuse* **then**
    r.end-time ← *min(alert1.end-time, alert2.end-time)*
    for *each attr:attribute except start-time, end-time, reference, alertid* **do**

r.attr ← *alert1.attr* ∪ *alert2.attr*
    end for
    fuse ← *false*
else
    if *thread* **then**
        r.end-time ← *max(alert1.end-time, alert2.end-time)*
        r.analyzer = *alert1.analyzer* ∪ *alert2.analyzer*
        thread ← *false*
    end if
end if
if *alert1.name* = *alert2.name* **then**
    r.name ← *alert1.name*
else
    r.name ← *"Attack Thread"*
end if
for *each attr:attribute except start-time, end-time, reference, analyzer, alertid* **do**
    if *alert1.attr* = *alert2.attr* **then**
        r.attr ← *alert1.attr*
    else
        r.attr ← **null**
    end if
end for
return *r*

end

In the *focus recognition* component, alerts are aggregated then classified using feature similarity. Unrelated and false alerts tend to be random and will not correlate, hence uncorrelated alerts are removed by *shushing the alerts* component. Lastly, *multi-step correlation*, is expected to achieve substantial improvement in the abstraction level and data reduction [13]. In this component, priori information of the network topology, known scenarios, etc are provided by the expert knowledge DB; hence high level patterns are specified.

In the *intention recognition* component, relevant behavior is grouped into attack scenarios to extract attack strategy and plan recognition.

In the final component, *impact analysis*, the asset DB is consulted to determine all services that are dependent on a specific target. The heartbeat monitor checks whether all dependent services are still operational. If any service is failed, this information can be added to the alert as a likely consequence of the attack [5].

The asset DB stores information about installed network services, dependencies among services, and their importance to the overall operation of a network installation. So the DB does not represent an absolute measure of the importance of any asset, but rather reflect the subjective view of a security administrator. It is updated if there is any new information from the impact analysis or prioritization components.

The *knowledge DB* is a complete repository including all the necessary information about attacks, vulnerabilities, and the topological and configuration information about the protected networks and hosts.

III. IMPLEMENTATION AND EXPERIMENTAL RESULTS

In order to evaluate the improved alert correlation model performance, a collection of 10 experiments on DARPA 2000 scenarios datasets and a testbed network dataset have been carried out. Each component's function of the innovative alert correlation framework is explained in details, together with the implementation of the model based on the improved

TABLE I: Impact of Preprocessing Component on LLDOS Scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	891	922	430	494
Output alerts	886	922	425	489

TABLE II: Impact of Prioritization Component on LLDOS Scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	886	922	425	489
Output alerts	188	167	54	71
Reduction Rate	78.78%	81.89%	87.29%	85.48%

TABLE III: Impact of Alert Fusion Component on LLDOS Scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	188	167	54	71
Output alerts	92	110	34	45
Reduction Rate	51.06%	34.13%	37.04%	36.62%

framework. In the implementation, Microsoft SQL Server 2005 was used as the relational database to store the alert datasets, the intermediate data, and the analysis results of each component as well as the correlated alerts. Programs written in C#, Microsoft Visual Studio 2010, were created to implement the correlation components' functionalities. The alert log files generated by RealSecure IDS of the DARPA simulation network is used [14] in eight experiments, which are explained in the next section.

A. Experiments on DARPA 2000 Datasets

DARPA 2000 [15] is a well-known IDS evaluation dataset created by the MIT Lincoln Laboratory. It consists of two multistage attack scenarios, namely Lincoln Laboratory DoS Data Sets Scenario (LLDOS) 1.0 and LLDOS 2.0.2. Each scenario includes network traffic collected from both the Demilitarized Zone (DMZ) and the inside part of the evaluation network. Eight experiments were performed, four on the improved model and four on the comprehensive approach model.

1) Data Normalization Unit:
- **Normalization**
  The alerts were already normalized, and in IDMEF standard format [8].
- **Preprocessing**
  In both scenarios, there are 45 features, of which only 7 features were extracted, namely EventID, timesec, SrcIPAddress, DestPort, DestIPAddress, OrigEventName, and SrcPort. The date attribute was represented in date/time format, and was then converted to time in seconds (represented as timesec). 5 alerts, representing incomplete data, were removed in all datasets, except for the inside segment of scenario 1.0., see Table I.

2) Filter-based Correlation Unit:
The primary goal is to reduce the number of alerts to be correlated by eliminating false, irrelevant and low risk alerts. False alerts need to be handled at an early stage as they will have negative impact on the correlation result, and moreover the number of processed alerts will be greatly reduced.

- **Prioritization**
  The ranking/priority of alerts of LLDOS scenarios from [16] is used. Thus low risk alerts are discarded, and only the medium and high risk alerts are sent to the next component. Table II shows the implementation results.

- **Alert Verification**
  This requires that the protected assets be available for real-time verification of the actual exposure of the system

and/or that a detailed model of the installed network services be available. Unfortunately, this information is not available for the data set analyzed and there is no sufficient information found about the asset DB, so this component could not be implemented.

3) Data Reduction Unit:
Similar alerts are fused and thus data is reduced by eliminating data redundancies, and irrelevant, false and unreal alarms using alert correlation. False alerts are usually less likely to be correlated using alert correlation.

- **Alert Fusion**
  There were two sensors in DARPA data sets, but all the alerts generated by one of the sensors contained null and incomplete values and thus were removed by the preprocessing component. Thus, there were no fusion in the data set used as all traffic injected into this component were seen by one sensor, but there were thread reconstruction. Table III shows the results of the implementation.

- **Focus Recognition**
  This component has the task of identifying hosts that are either the source or the target of a substantial number of attacks. This is used to identify Denial-of-Service (DoS) attacks or port scanning attempts. It aggregates the alerts associated with single hosts attacking multiple victims (called a one-to-many scenario) and single victims that are targeted by multiple attackers (called a many-to-one scenario).
  The one-to-many scenario has two tunable parameters: the size of the timeout, which is used for the initial window size, and the minimum number of alerts for a meta-alert to be generated. On the other hand, the many-to-one scenario has three tunable parameters: the first two are the same as for the one-to-many scenario. The third parameter is the number of meta-alerts required before a many-to-one alert is labeled as a denial-of-service attack [5][6][17].
  In the carried out experiments, the minimum number of alerts in a meta-alert was two. We first applied one-

TABLE IV: Impact of Focus Recognition Component (one-to-many)

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	92	110	34	45
Output alerts	42	57	23	27
Reduction Rate	56.52%	48.18%	32.35%	40%

TABLE V: Impact of Focus Recognition Component (many-to-one)

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	42	57	23	27
Output alerts	31	28	5	24
Reduction Rate	26.19%	50.88%	78.26%	11.11%

TABLE VI: Impact of Shushing the alerts Component on LLDOS Scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	31	28	5	24
Output alerts	6	7	3	5
Reduction Rate	80.65%	75%	40%	79.17%

TABLE VII: Impact of Multi-step Correlation Component on LLDOS Scenarios

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	6	7	3	5
Output alerts	6	5	2	4
Reduction Rate	0%	28.57%	33.33%	20%

to-many focus recognition on DARPA datasets, then followed by many-to-one focus recognition. Some horizontal scan and multi-scan attacks were observed. Tables IV and V show the reduction rates. DMZ in scenario 2.0.2 shows a great RR as is expected being a multistage attack scenario.

- **Shushing the Alerts**
  As shown in [18], alert correlation can be used to differentiate between false and true alerts. False alerts and unreal alarms tend to be more random than actual alerts, and are less likely to be correlated. Thus, based on this founding, we intentionally removed the alerts that are not correlated in the alert fusion and focus recognition, resulting in Table VI, which shows great reduction rates.

- **Multi-step Correlation**
  The goal of this component is to identify high-level attack patterns that are composed of several individual attacks. The high-level patterns are usually specified using some form of expert knowledge [2][5][17][19].
  Relying on the information in [20], attack patterns are identified, and used to implement this component resulting in Table VII.

4) Intention Recognition:
Intention or plan recognition is the process of inferring

the goals of an intruder by observing his/her actions [21]. It deduces strategies and objectives of attackers based on attack scenarios that are output by correlation systems. Failed attacks can be useful to know so to be avoided in the future. Using alert correlation, the intruders' relevant behavior can be grouped into attack scenarios, and later on, their attack strategy or plan can be extracted and fed back to update the expert knowledge DB.

Inadequate information of attack strategies or plans of intruders in the data set used hindered the implementation of this component.

5) Impact Analysis:
This component contextualizes the alerts with respect to a specific target network. It combines the alerts from the previous correlation components with data from an asset DB and a number of heartbeat monitors to determine the impact of the detected attacks on the operation of the monitored network and on the assets that are targeted by the attacker. Thus, it requires a precise modeling of the relationships among assets in a protected network and constant health monitoring of those assets. Hence, insufficient information of asset DB of LLDOS scenarios deters the implementation.

B. Experiments on Testbed Network

Two Snort 2.9.3.1 IDSs were installed in a Linux machine, and a Windows XP machine. Attacks were launched remotely and the alert log files were analyzed. Two experiments were performed. Tables VIII and IX show the reduction rates using the improved approach were 96.1% compared to 94.81% using the comprehensive approach.

C. Summary of Experiments on DARPA 2000 Scenarios

Tables X and XI displays the number of processed alerts and the total reduction rates using the novel approach and the Comprehensive approach respectively on each of the LLDOS scenarios 1.0 and 2.0.2.

Table XII presents a summary of the total alert reduction for each dataset. Fig. 2 illustrates the effect of the improved correlation model on LLDOS 1.0 and 2.0.2 scenarios. There is a substantial drop in the number of alerts in the priority component for all datasets. Since the processing time is proportional to the number of processed alerts, hence both Fig. 2 and 3 assured the affirmation of the better performance of the novel improved model over the Comprehensive approach.

1) Comparison of the Performance of the Improved Model with the Comprehensive Approach on LLDOS Scenario 1.0:
Tables XIII and XIV show the number of processed alerts in each component for scenario 1.0 for the improved model compared to the Comprehensive approach discussed in [5]. Since the processing time is proportional to the number of processed alerts, hence Fig. 4 shows that the improved model gives better results.

TABLE VIII: No. of Processed Alerts and the total RR using Improved Model for the Testbed Network

	Prepr.	Prio.	Fus.	1:M	M:1	Shush.	Multi	total
No. of Alerts	77	44	16	9	6	5	3	159
Reduction Rate	0%	42.86%	63.63%	43.75%	33.33%	16.67%	40%	96.1%

TABLE IX: No. of Processed Alerts and the total RR using Comprehensive Approach for the Testbed Network

	Prepr.	Fus.	1:M	M:1	Multi	Prio.	total
No. of Alerts	77	40	15	13	12	4	161
Reduction Rate	0%	48.05%	62.5%	13.33%	7.69%	66.67%	94.81%

TABLE X: No. of Processed Alerts and the Total Reduction Rates (RR) using Novel Model for LLDOS Scenarios

	Prepr.	Prio.	Fus.	1:M	M:1	Shush.	Multi	# of Processed Alerts	RR
DMZ 1.0	886	188	92	42	31	6	6	1251	99.33.%
Inside 1.0	922	167	110	57	28	7	5	1296	99.46%
DMZ 2.0.2	425	54	34	23	5	3	2	546	99.53%
Inside 2.0.2	489	71	45	27	24	5	4	665	99.19%

TABLE XI: No. of Processed Alerts and the Total Reduction Rates (RR) using Comprehensive Approach for LLDOS Scenarios

	Prepr.	Fus.	1:M	M:1	Multi.	Prio.	# of Processed Alerts	RR
DMZ 1.0	886	619	208	175	118	13	2019	98.54%
Inside 1.0	922	622	193	151	107	16	2011	98.26%
DMZ 2.0.2	425	241	63	46	44	5	824	98.84%
Inside 2.0.2	489	276	71	44	33	7	920	98.57%
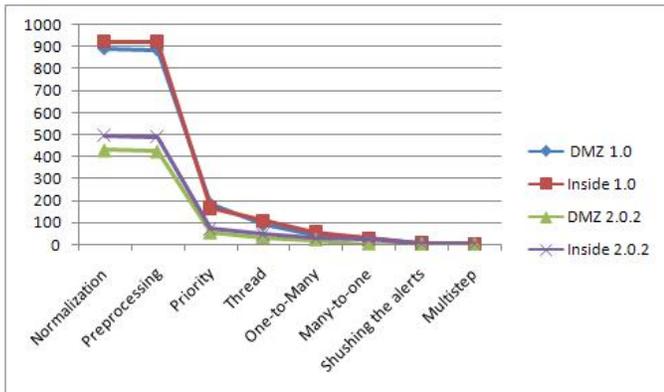
Fig. 2: Effect of Improved Correlation Model on LLDOS Scenarios 1.0 and 2.0.2.
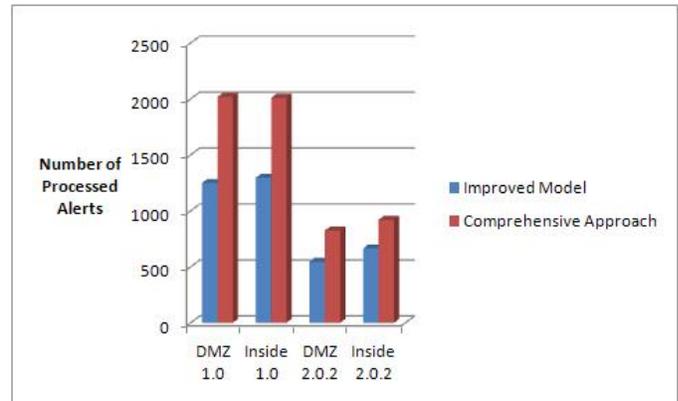
TABLE XII: Total Alert Reduction for the Improved Model

	DMZ 1.0	Inside 1.0	DMZ 2.0.2	Inside 2.0.2
Input alerts	891	922	430	494
Output alerts	6	5	2	4
Reduction Rate	99.33.%	99.46%	99.53%	99.19%

TABLE XIII: No. of Processed Alerts using Improved Model for Scenario 1.0

|	Prepr	Prio	Fus	1:M	M:1	Shush	Multi	total
DMZ	886	188	92	42	31	6	6	1251
Inside	922	167	110	57	28	7	5	1296

Fig. 3: Comparison of Processing Time of Improved Correlation Model and Comprehensive Approach on LLDOS Scenarios 1.0. and 2.0.2

TABLE XIV: No. of Processed Alerts using Comprehensive Approach for Scenario 1.0

	Prepr.	Fus.	1:M	M:1	Multi.	Prio.	total
DMZ	886	619	208	175	118	13	2019
Inside	922	622	193	151	107	16	2011

2) Comparison of the Performance of the Improved Model with the Comprehensive Approach on LLDOS Scenario 2.0.2: Tables XV and XVI show the number of processed alerts in each component for scenario 2.0.2 for the improved model and the Comprehensive approach [5] respectively. The graph

TABLE XV: No. of Processed Alerts using Improved Model for Scenario 2.0.2

	Prepr	Prio	Fus	1:M	M:1	Shush	Multi	total
DMZ	425	54	34	23	5	3	2	546
Inside	489	71	45	27	24	5	4	665

TABLE XVI: No. of Processed Alerts using Comprehensive Approach for Scenario 2.0.2

	Prepr.	Fus.	1:M	M:1	Multi.	Prio.	total
DMZ	425	241	63	46	44	5	824
Inside	489	276	71	44	33	7	920

of Fig. 5 assured the affirmation of the better performance of the innovative improved model.

D. Summary of Experiments on Testbed Network

1) Testing the Improved model on a Testbed Network:
The functionality of the improved alert correlation model was validated by processing this dataset and hence verifying that the attack patterns were recognized.

2) Comparing the Improved Model with the Comprehensive Approach using the Testbed Network:
Fig. 6 shows that the total number of processed alerts using the improved approach is 159 compared to 161 using the comprehensive approach respectively. The improved approach shows better results, although slight difference, but confirms the effectiveness of the improved alert correlation model.

IV. RELATED WORK

Valeur et al in [5] presented a complete comprehensive set of components. Their experiments demonstrated that the effectiveness of each component is dependent on the data sets being analyzed, and each component can contribute to the overall performance.

Yu et al. presents a collaborative architecture for multiple IDSs to detect real-time network intrusions.The architecture is composed of three parts: Collaborative Alert Aggregation, Knowledge-based Alert Evaluation and Alert Correlation to cluster and merge alerts from multiple IDS products to achieve an indirect collaboration among them [22].

Also Depren et al. proposed a novel IIDS architecture utilizing both anomaly and misuse detection approaches, together with a decision support system to combine their results [23]. In the same year, Zhang et al. suggested a distributed IDS based on Clustering with unlabeled data [24]. Later in that year, Katti et al. presented the first wide-scale study of correlated attacks, and their results showed that collaborating IDSs need to exchange alert information in realtime [25].

Sadoddin and Ghorbani showed an overall view of the applied techniques which have been used for different aspects of correlation. The techniques were presented in the context of a comprehensive correlation framework. As high-level comparison between techniques, either competitive or
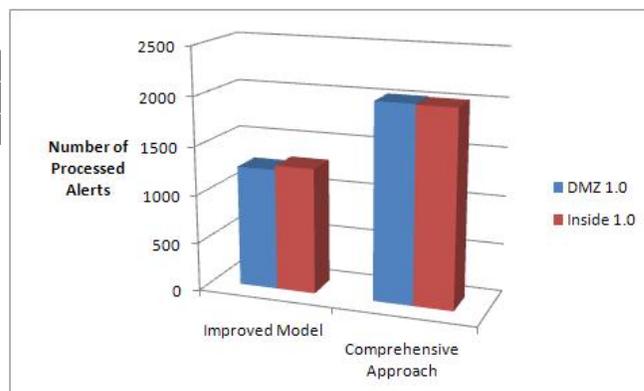

Fig. 4: Comparison of Processing Time of Improved Correlation Model and Comprehensive Approach on LLDOS Scenario 1.0.
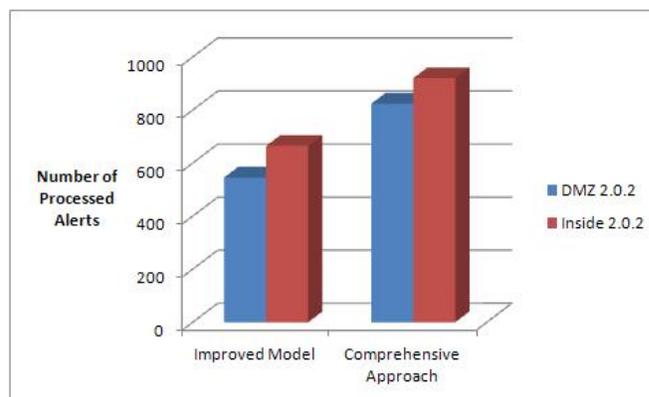

Fig. 5: Comparison of Processing Time of Improved Correlation Model and Comprehensive Approach on LLDOS Scenario 2.0.2.
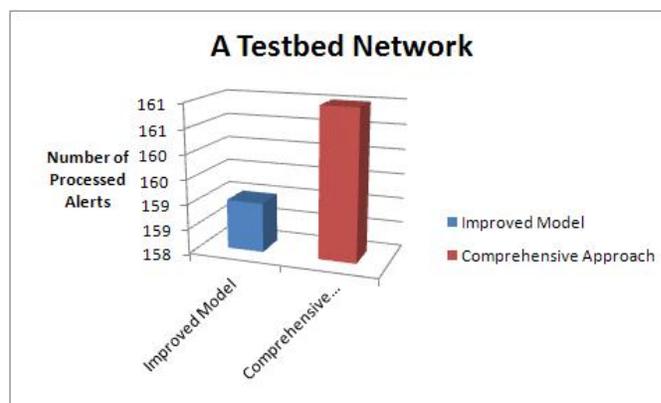

Fig. 6: Comparison of Processing Time of Improved Correlation Model and Comprehensive Approach on the Testbed Network

complementary to each, the pros and cons of the techniques were described from their point of view [26].

From the analysis in [13], researchers propose an improved solution for an alert correlation technique based on six capabilities criteria identified which are capabilities to perform alert reduction, alert clustering, identify multi-step attacks, reduce false alert, and to detect known and unknown attacks.

In February 2009, Zhou et al. proposed a decentralized, multi-dimensional alert correlation algorithm for CIDSs. A two-stage algorithm, implemented in a fully distributed CIDS, first clusters alerts locally at each IDS, before reporting significant alert patterns to a global correlation stage [3]. Later in the same year, Zhou et al. proposed a decentralized, multi-dimensional alert correlation algorithm for CIDSs. A two-stage algorithm, implemented in a fully distributed CIDS, first clusters alerts locally at each IDS, before reporting significant alert patterns to a global correlation stage. They summarized the current research directions in detecting coordinated attacks using CIDSs. In particular, two main challenges in CIDS research: CIDS architectures and alert correlation algorithms are highlighted and analyzed [1].

Sadoddin and Ghorbani proposed a framework for real-time alert correlation which incorporates novel techniques for aggregating alerts into structured patterns and incremental mining of frequent structured patterns [27].

In [28], Taha et al presented an agent-based alert correlation model. A learning agent learns the nature of dataset to select which components to be used and in which order. They proved that their method achieved minimum alerts to be processed on each component, depending on the dataset, and minimum time for correlation process. Their method differs from ours, in that they have learning agent, and we specify an order of the components which gives better performance by processing less number of alerts, hence minimum correlation time as only the high risk alerts are processed.

Ghorbani et al in [21] showed an overall view of the applied techniques which have been used for different components of an alert correlation framework.

Meinel et al in [29] identified the data storage and processing algorithms to be the most important factors influencing the performance of clustering and correlation. They proposed and implemented the utilization of memory-supported algorithms and a column-oriented DB for correlation and clustering in an extensible IDS correlation platform.

In October 2011, an alert correlation architecture is proposed by Amiri et al. Their architecture consists of four important components namely: log management, alert correlation, incident response and knowledge base system. The proposed architecture uses anomaly-based analysis in the alert correlation component. They reviewed and compared different techniques for alert correlation. Their study finally proposes that a hybrid model of multiple techniques leads to better performance of alert correlation engine [30].

Early the following year, in April 2012, Njogu et al. proposed a comprehensive approach to address the shortcomings of the vulnerability based alert management approaches. They

proposed a fast and efficient approach that improves the quality of alerts as well as reduce the volumes of redundant alerts generated by signature based IDSs. Their approach has several components that are presented in three stages: Stage 1 involves alert pre-processing, correlation of alerts against the meta alert history and verification of alerts against Enhanced Vulnerability Assessment (EVA) data; Stage 2 involves classification of alerts based on their alert metrics; and Stage 3 involves correlation of alerts in order to reduce the redundant and isolated alerts as well discover the causal relationships in alerts [31].

In the same month, Soleimani and Ghorbani took a different view and consider alert correlation as the problem of inferring an intruder's actions as alert patterns that are constructed progressively. Their work is based on a multi-layer episode mining and filtering algorithm. A decision-tree-based method is used for learning specifications of each attack pattern and detecting them in alert streams. They also used a Correlation Weight Matrix (CWM) for encoding correlation strength between attack types in the attack scenarios. One of the distinguishing features of their proposed technique is detecting novel multi-step attack scenarios, using a rule prediction method. The results have shown that their approach can effectively discover known and unknown attack strategies with high accuracy. They actually achieved more than 90% reduction in the number of discovered patterns while more than 95% of final patterns were actual patterns. Furthermore, their rule prediction capability showed a precise forecasting ability in guessing future alerts [32].

In July 2012, Amaral et al. presented an automated alarm correlation system composed of three layers, which obtains raw alarms and presents to network administrator a wide view of the scenario affected by the volume anomaly. In the preprocessing layer, the alarm compression is performed using their spatial and temporal attributes, which are reduced into a unique alarm named Device Level Alarm (DLA). The correlation layer aims to infer the anomaly propagation path and its origin and destination using DLAs and network topology information. The presentation layer provides the visualization of the path and network elements affected by the anomaly propagation. Moreover, the Anomaly Propagation View (APV) is presented, which is a graphical tool developed to provide a wide visualization of the network status [33].

Lately in September of the same year, Mohamed et al. constructed a holistic solution that is able to reduce the number of alerts to be processed and at the same time produced a high quality attack scenarios that are meaningful to the administrators in a timely manner. Their proposed framework and the novel clustering method, architectured solely with the intention of reducing the amount of alerts generated by IDS. The clustering method was tested against two datasets; a globally used dataset, DARPA and a live dataset from a cyber attack monitoring unit that uses Snort engine to capture the alerts [34].

V. Conclusion and Future Work

This paper presents an innovative alert correlation framework based on a Collaborative Intelligent Intrusion Detection System (CIIDS) architecture. Alert correlation analyzes the alerts and aims to relate different alerts to build a big picture of the attack, thus giving a high-level view of the security status. The innovative framework attempts to minimize the number of processed alerts on each component and thus minimizing the correlation processing time. Hence, the correlation model components are reordered in such a way that achieves better performance by processing less number of alerts. It removes irrelevant, unreal and false alerts in the early phases of the correlation by reordering the components. Uncorrelated alerts are also dealt with in a new component, *shushing the alerts* in order to discard irrelevant and false positives. Any alert that is not correlated after being processed by a number of components is deliberately removed. An algorithm for this new component is presented. The performance is improved after the attention is focused on correlating higher severity alerts. High level patterns are specified in the multi-step component. The impact of the attack on the network assets and services is also investigated. Thus by diverting more resources to deal with high risk/priority alerts to be correlated, the effectiveness of alert correlation is significantly improved.

Further experiments and comparisons with different datasets and a real network dataset will be investigated. Several research directions exist. The generation of datasets to be utilized in evaluations is essential. Furthermore, there is a need for a mission model and its relationship to the network assets, and also a health monitoring system to determine the impact of the attacks on the network. Another research direction is the investigation of applying soft computing techniques to enhance the correlation process.

Acknowledgment

References

[1] C. V. Zhou, C. Leckie and S. Karunasekera, *A survey of coordinated attacks and collaborative intrusion detection.* Elsevier Ltd., Computer Security, pp. 1–17, June 2009.

[2] H. T. Elshoush and I. M. Osman, *Alert Correlation in Collaborative Intelligent Intrusion Detection Systems - A Survey.* Elsevier Ltd., Journal of Applied Soft Computing, **11**, Issue No. 7, 4349–4365, October 2011.

[3] C. V. Zhou, C. Leckie and S. Karunasekera, *Decentralized multidimensional alert correlation for collaborative intrusion detection.* Elsevier Ltd., Journal of Network and Computer Applications, **32**, 1106–1123, Feb. 2009.

[4] R. Bye, S. A. Camtepe and S. Albayrak, *Collaborative Intrusion Detection Framework: Characteristics, Adversarial Opportunities and Countermeasures*, August 2010.

[5] F. Valeur, G. Vigna, C. Kruegel and R. Kemmerer, *A Comprehensive Approach to Intrusion Detection Alert Correlation.* IEEE Transactions on Dependable and Secure Computing, the IEEE Computer Society, **1**, Issue No. 3, 146–169, July-September, 2004.

[6] F. Valeur, *Real-time ID Alert Correlation.* PhD Thesis, June 2006.

[7] H. T. Elshoush and I. M. Osman, *An Improved Framework for Intrusion Alert Correlation.* Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2012, WCE 2012, 4-6 July, 2012, London, U.K., pp. 518–523.

[8] H. Debar, D. Curry and B. Feinstein, *The Intrusion Detection Message Exchange Format (IDMEF).*, 2007. http://www.ietf.org/rfc/rfc4765.txt.

[9] A. Zainal1, M. A. Maarof1 and S. M. Shamsuddin, *Features Selection Using Rough-PSO in Anomaly Intrusion Detection*, 2007.

[10] A. Zainal, M. A. Maarof and S. M. Shamsuddin, *Feature Selection Using Rough Set in Intrusion Detection.*

[11] F. Amiri, M. M. R. Yousefi, C. Lucas and A. Shakery, *Improved Feature Selection for Intrusion Detection System.* Elsevier Ltd., Journal of Network and Computer Applications, Jan. 2011.

[12] J. J. Davis and A. J. Clark, *Data Preprocessing for Anomaly-based Network Intrusion Detection: A Review.* Elsevier Ltd., Journal of Computer and Security, 353–375, October 2011.

[13] R. Yusof, S. R. Selamat and S. Sahib, *Intrusion Alert Correlation Technique Analysis for Heterogeneous Log.* International Journal of Computer Science and Network Security (IJCSNS), **8** Issue No. 9, 132–138, Sept. 2008.

[14] P. Ning, *TIAA: A Toolkit for Intrusion Alert Analysis*, 2007. http://discovery.csc.ncsu.edu/software/correlator/.

[15] MIT Lincoln Laboratory 2000 DARPA Intrusion Detection Scenario Specific Datasets. http://www.ll.mit.edu/index.html.

[16] M. M. Siraj, M. A. Maarof and S. Z. M. Hashim, *Intelligent Alert Clustering Model for Network Intrusion Analysis.* Published by ICSRS Publication, Int. J. Advance. Soft Comput. Appl., **1**, Issue No. 1, July 2009, ISSN 2074-8523.

[17] C. Kruegel, F. Valeur and G. Vigna, *Intrusion Detection and Correlation - Challenges and Solutions.* Springer, a textbook, 2005.

[18] P. Ning,, Y. Cui and D. S. Reeves, *Constructing Attack Scenarios through Correlation of Intrusion Alerts.* In Proceedings of the 9th ACM Conference on Computer and Communications Security, Washington D.C., pp. 245–254, November 2002.

[19] P. Ning, Y. Cui and D. S. Reeves, *Analyzing Intensive Intrusion Alerts Via Correlation.* In Proceedings of the 5th International Symposium on Recent Advances in Intrusion Detection (RAID 2002), LNCS 2516, Zurich, Switzerland, pp. 74–94, October 2002.

[20] Y. Cui, *A Toolkit for Intrusion Alerts Correlation Based on Prerequistes and Consequences of Attacks.* MSc. Thesis, Dec. 2002.

[21] A. A. Ghorbani, W. Lu and M. Tavallaee, *Network Intrusion Detection and Prevention: Concepts and Techniques.* Springer, a textbook, 2010.

[22] J. Yu, Y. V. R. Reddy, S. Selliah, S. Reddy, V. Bharadwaj, and S. Kankanahalli, *Trinetr: An architecture for collaborative intrusion detection and knowledge-based alert evaluation.* Elsevier Ltd. Advanced Engineering Informatics, Vol. 19:pp. 93–101, May 2005.

[23] O. Depren, M. Topallar, E. Anarim, and M. K. Ciliz, *An intelligent intrusion detection system (ids) for anomaly and misuse detection in computer networks.* Elsevier Ltd. Expert Systems with Applications, Vol. 29:pp. 713–722, May 2005.

[24] Y. F. Zhang, Z. Y. Xiong and X. Q. Wang, *Distributed intrusion detection based on clustring.* Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou, August 2005.

[25] S. Katti, B. Krishnamurthy and D. Katabi, *Collaborating against common enemies.* USENIX Association. Internet Measurement Conference, pages 365–378, October 2005.

[26] R. Sadoddin and A. Ghorbani, *Alert correlation survey: Framework and techniques.* ACM, Markham, Ontario, Canada, Oct-Nov. 2006.

[27] R. Sadoddin and A. A. Ghorbani *An incremental frequent structure mining framework for real-time alert correlation.* Elsevier Ltd. Computers Security, Vol. 28:pp. 153–173, 2009.

[28] A. E. Taha, I. Abdel Ghaffar, A. M. Bahaa Eldin and H. M. K. Mahdi, *Agent Based Correlation Model For Intrusion Detection Alerts.* IEEE Computer Society, May 2010.

[29] S. Roschke, F. Cheng and C. Meinel, *A Flexible and Efficient Alert Correlation Platform for Distributed IDS.* Fourth International Conference on Network and System Security, 2010.

[30] F. Amiri, H. Gharaee and A. R. Enayati, *A complete operational architecture of alert correlation.* International Conference on Computational Aspects of Social Networks (CASoN), October 2011.

[31] H. W. Njogu, L. Jiawei, J. N. Kiere and D. Hanyurwimfura, *A comprehensive vulnerability based alert management approach for large networks.* Elsevier Ltd. Future Generation Computer Systems, April 2012.

[32] M. Soleimani and A.A. Ghorbani, *Multi-layer episode filtering for the multi-step attack detection.* Computer Communications, doi:10.1016/j.comcom.2012.04.001, April 2012.

[33] A. A. Amaral, B. B. Zarpelao, L. de S. Mendes, J. J. P. C. Rodrigues and M. L. Proenca Junior, *Inference of network anomaly propagation using spatio-temporal correlation.* Journal of Network and Computer Applications, July 2012.

[34] A. B. Mohamed, N. B. Idris and B. Shanmugum, *An operational framework for alert correlation using a novel clustering approach.* International Journal of Computer Applications (0975 8887), Volume 54, Issue No.12, September 2012.