





Jan Zizka  
Dhinaharan Nagamalai (Eds)

# Computer Science & Information Technology

Sixth International conference on Computer Science and Information  
Technology (CCSIT 2016)  
Zurich, Switzerland, January 02~03, 2016



**AIRCC Publishing Corporation**

## **Volume Editors**

Jan Zizka,  
Mendel University in Brno, Czech Republic  
E-mail: zizka.jan@gmail.com

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

ISSN: 2231 - 5403  
ISBN: 978-1-921987-45-8  
DOI : 10.5121/csit.2016.60101 - 10.5121/csit.2016.60127

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

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

## Preface

The Sixth International conference on Computer Science and Information Technology (CCSIT 2016) was held in Zurich, Switzerland, during January 02~03, 2016. The Fourth International Conference on Signal, Image Processing and Pattern Recognition (SIPP 2016), The Fourth International Conference on Artificial Intelligence, Soft Computing (AISC 2016), The Fourth International Conference on Control, Modelling, Computing and Applications (CMCA 2016), The Fourth International Conference on Software Engineering and Applications (SEAS-2016), The International Conference on Computer Science, Information Technology (CSITEC 2016), The International Conference on Data Mining & Knowledge Management (DaKM 2016), The Fifth International Conference on Parallel, Distributed Computing Technologies and Applications (PDCTA 2016) and The Seventh International Conference on Networks & Communications (NeCoM 2016) were collocated with the CCSIT-2016. The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

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

The CCSIT-2016, SIPP-2016, AISC-2016, CMCA-2016, SEAS-2016, CSITEC-2016, DaKM-2016, PDCTA-2016, NeCoM-2016 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer network and communications research.

In closing, CCSIT-2016, SIPP-2016, AISC-2016, CMCA-2016, SEAS-2016, CSITEC-2016, DaKM-2016, PDCTA-2016, NeCoM-2016 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the CCSIT-2016, SIPP-2016, AISC-2016, CMCA-2016, SEAS-2016, CSITEC-2016, DaKM-2016, PDCTA-2016, NeCoM-2016.

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

Jan Zizka  
Dhinaharan Nagamalai

## Organization

### General Chair

Jan Zizka  
Dhinaharan Nagamalai

Mendel University in Brno, Czech Republic  
Wireilla Net Solutions PTY LTD, Australia

### Program Committee Members

Abd El-Aziz Ahmed  
Abdelkrim Haqiq  
Abdolreza Hatamlou  
Aftab Alam  
Ahmed H.Aliwy  
Aleksandar Sugaris  
Ali Asghar Safaei  
Ali Elkateeb  
Amelia Zafra Gomez  
Ankit Chaudhary  
Asadullah Shaikh  
Assem Abdel Hamied Moussa  
Ayad Salhieh  
Chen ZhiQiang  
Chiranjib Sur  
Dana Petcu  
Dongchen Li  
Elashiry M.A  
Epaminondas Kapetanios  
Ercan Oztemel  
Faiz ul haque Zeya  
Farzad Kiani  
Hamid Mcheick  
Hassan Chizari  
Héldon José Oliveira Albuquerque  
Hossein Jadidoleslamy  
Huiyu Zhou  
Isa Maleki  
Jan Lindström  
Jianfeng Wang  
José Raniery  
Kassim S.Mwitondi  
Kemal  
Kun Ren  
Kwangjin Park  
Lin Wang  
Luis Fernando de Mingo Lapez  
Maher Ben Jemma

Cairo University, Egypt  
Hassan 1st University, Morocco  
Islamic Azad University, Iran  
King Khalid University, Saudi Arabia  
University of Kufa, Iraq  
ICT College, Serbia  
Tarbiat Modares University, Iran  
University of Michigan-Dearborn, USA  
University of Cordoba, Spain  
Truman State University, USA  
Najran University, Saudi Arabia  
Cairo University, Egypt  
Australian College of Kuwait, Kuwait  
University of Missouri-Kansas City, USA  
University of Florida, US  
West University of Timisoara, Romania  
Peking University, China  
Beni Suef University, Egypt  
University of Westminster, London  
Marmara University, Turkey  
Bahria University, Pakistan  
Istanbul S.Zaim University, Turkey  
University of Quebec at Chicoutimi, Canada  
University Technology Malaysia, Malaysia  
Integrated Faculties of Patos (FIP), Brazil  
MUT University, Iran  
Queen's University Belfast, United Kingdom  
Islamic Azad University, Iran  
MariaDB Corporation, Finland  
Xidian University, China  
University of São Paulo, Brazil  
Sheffield Hallam University, United Kingdom  
Abant Izzet Baysal University, Turkey  
Yale University, United States  
Wonkwang University, South Korea  
University of Jinan, China  
Technical University of Madrid, Spain  
University of Sfax, Tunisia

Marc Sevaux	Universite de Bretagne, France
Marta Beltran Pardo	Universidad Rey Juan Carlos, Spain
Maryam Rastgarpour	Islamic Azad University, Iran
Marystella Amaldas	Saigon International University, Vietnam
Mayyash	The California State University(CSU), USA
Mehrdad Jalali	Mashhad Azad University, Iran
Mohammed Amin	Higher Colleges of Technology, UAE
Moses Ekpenyong	University of Uyo, Nigeria
Muhammad Sarfraz	Kuwait University, Kuwait
Muhammad Shafie bin Abd Latiff	Universiti Teknologi Malaysia, Malaysia
Narasimha I.V.	University of Houston, USA
Natarajan Meghanathan	Jackson State University, USA
Nouh Sabri	Cairo University, Egypt
Odey J.A	Federal University Wukari, Nigeria
Othmane Alaoui Fdili	Mohammed V University, Morocco
Ouksel M	The University of Illinois at Chicago, Illinois
Patrizia Scandurra	University of Bergamo, Italy
Pelin Angin	Purdue University, USA
Peter Ogedebe	BAZE University, Nigeria
Piet Kommers	University of Twente bld Cubicus, Netherlands
Raed Ibraheem Hamed	University of Human Development, Iraq
Rafah M. Almuttairi	University of Babylon, Iraq
Ramayah T	Universiti Sains Malaysia, Malaysia
Rangiha Mohammad	City University London, UK
Rastislav Roka	Slovak University of Technology, Slovakia
Rauhi Ibrahim Elkhatib	Thamar University, Yemen
Renuka Mohanraaj	Maharishi University of Management, USA
Resen Sallama	Erciyes University, Turkey
Ricardo De Carvalho Destro	University Center of FEI, Brazil
Rim Haddad	Innov'com Laboratory, Tunisia
Rosziati Ibrahim	Universiti Tun Hussein Onn Malaysia, Malaysia
Saeid Asgari Taghanaki	Azad University, Iran
Salama Zoiny	Erciyes University, Turkey
Samadhiya	National Chiao Tung University, Taiwan
Sebastian Ventura	University of Cordoba, Spain
Seyed Ziaeddin Alborzi	Université de Lorraine, France
Shakir Khan	Leading University, Bangladesh
Shigeru Yamada	Tottori University, Japan
Sokyna Qatawneh	Alzaytoonah University of Jordan, Jordan
Subidh Ali	New York University Abu Dhabi, UAE
Tanweer Alam	Islamic University, Kingdom of Saudia Arabia
Terumasa Aoki	Tohoku University, Japan
Vishal Zinjuvadia	Dell, United States
Waldir Sabino	Federal University of Amazonas, Brazil
Weifa Liang	Australian National University, Australia
Yassine Boukal	University of Lorraine, France
Yong-Jin Lee	Korea National University of Education, Korea
Youssef Fakhri	University Ibn Tofail, Morocco
Zaher Al Aghbari	University of Sharjah, UAE

## **Technically Sponsored by**

**Networks & Communications Community (NCC)**



**Computer Science & Information Technology Community (CSITC)**



**Digital Signal & Image Processing Community (DSIPC)**



## **Organized By**



**Academy & Industry Research Collaboration Center (AIRCC)**

## TABLE OF CONTENTS

### **The Sixth International conference on Computer Science and Information Technology (CCSIT 2016)**

<b>Infrastructure Consolidation for Interconnected Services in a Smart City Using Cloud Environment.....</b>	01 - 12
<i>Jorge F Hernandez, Victor M Larios, Manuel Avalos and Ignacio Silva-Lepe</i>	
<b>Similarity Analysis of DNA Sequences Based on the Chemical Properties of Nucleotide Bases, Frequency and Position of Group Mutations.....</b>	13 - 22
<i>Fatima KABLI, Reda Mohamed HAMOU and Abdelmalek AMINE</i>	
<b>About the Suitability of Clouds in High-Performance Computing.....</b>	23 - 33
<i>Harald Richter</i>	
<b>Reference Architecture for SMAC Solutions.....</b>	35 - 43
<i>Shankar Kambhampaty and Sasirekha Kambhampaty</i>	
<b>Competence Building Framework Requirements for Information Technology for Educational Management.....</b>	45 - 51
<i>Rakeshh Mohan Bhatt</i>	
<b>Stable Marriage Problem with Ties and Incomplete Bounded Length Preference List Under Social Stability.....</b>	53 - 62
<i>Ashish Shrivastava and C. Pandu Ranga</i>	

### **The Fourth International Conference on Signal, Image Processing and Pattern Recognition (SIPP 2016)**

<b>Family of 2-Simplex Cognitive Tools and their Applications for Decision-Making and its Justifications.....</b>	63 - 76
<i>Yankovskaya Anna and Yamshanov Artem</i>	
<b>Sequential Clustering-Based Event Detection for Nonintrusive Load Monitoring.....</b>	77 - 85
<i>Karim Said Barsim and Bin Yang</i>	
<b>HOL, GDCT and LDCT for Pedestrian Detection.....</b>	87 - 96
<i>Sanaa Tayb, Youssef Azdoud, Aouatif Amine, Bouchra Nassih, Hanaa Hachimi and Nabil Hmina</i>	

## **The Fourth International Conference on Artificial Intelligence, Soft Computing (AISC 2016)**

- An Advanced Tool for Managing Fuzzy Complex Temporal Information.....** 97 - 116  
*Aymen Gammoudi, Allel Hadjali and Boutheina Ben Yaghlane*
- Reliability Evaluation of Software Architecture Styles.....** 117 - 129  
*Gholamreza Shahmohammadi*
- A Model Based on Sentiments Analysis for Stock Exchange Prediction – Case Study of PETR4, PETROBRAS, Brazil.....** 131 - 139  
*Milson L. Lima, Sofiane Labidi, Thiago P. do Nascimento, Nadson S. Timbó, Gilberto N. Neto and Marcus Vinicius Lima Batista*
- JPL : Implementation of a Prolog System Supporting Incremental Tabulation.....** 323 - 338  
*Taher Ali Ziad Najem and Mohd Sapiyan*

## **The Fourth International Conference on Control, Modelling, Computing and Applications (CMCA 2016)**

- Optimal Beam Steering Angles of a Sensor Array for a Multiple Source Scenario.....** 141 - 150  
*Sanghyouk Choi, Joohwan Chun, Inchan Paek and Jonghun Jang*
- A Switched-Antenna Nadir-Looking Interferometric SAR Altimeter for Terrain-Aided Navigation.....** 151 - 157  
*Inchan Paek, Jonghun Jang, Joohwan Chun and Jinbae Suh*
- Optimization in Engine Design via Formal Concept Analysis Using Negative Attributes.....** 159 - 172  
*Rodríguez-Jiménez, J. M, Cordero, P, Enciso, M and Mora, A*
- Combined Classifiers for Time Series Shapelets.....** 173 - 182  
*Ivan S. Mitzev and Nickolas H. Younan*
- Resilient Interface Design for Safety-Critical Embedded Automotive Software.....** 183 - 199  
*Harald Sporer, Georg Macher, Christian Kreiner and Eugen Brenner*

## **The Fourth International Conference on Software Engineering and Applications (SEAS-2016)**

**Near-Real-Time Parallel ETL+Q for Automatic Scalability in Bigdata**..... 201 - 218  
*Pedro Martins, Maryam Abbasi and Pedro Furtado*

**The Perceptions of Agile Methodology in South Africa**..... 219 - 227  
*Thierry Mbah Mbelli and Jainesh Jaintylal Hira*

**Advanced Cloud Privacy Threat Modeling**..... 229 - 239  
*Ali Gholami and Erwin Laure*

**A Taxonomy for Tools, Processes and Languages in Automotive Software Engineering**..... 241 - 256  
*Florian Bock, Daniel Homm, Sebastian Siegl and Reinhard German*

## **The International Conference on Computer Science, Information Technology (CSITEC 2016)**

**Application of Biclustering Technique in Machine Monitoring**..... 257 - 266  
*Marcin Michalak*

**Virtual Scene Construction of Large-Scale Cultural Heritage : A Framework Initiated from the Case Study of the Grand Canal of China**..... 267 - 286  
*Jian Tan and Shenghua Wang*

## **The International Conference on Data Mining & Knowledge Management (DaKM 2016)**

**A Prefixed-Itemset-Based Improvement for Apriori Algorithm**..... 287 - 296  
*Yu Shoujian and Zhou Yiyang*

## **The Fifth International Conference on Parallel, Distributed Computing Technologies and Applications (PDCTA 2016)**

**State Space Generation Framework Based on Binary Decision Diagram for Distributed Explicit Model Checking** ..... 297 - 305  
*Nacer Tabib, Jean Michel Ilie and Djamel Eddine Saidouni*

**The Seventh International Conference on Networks & Communications  
(NeCoM 2016)**

**A New Algorithm for Construction of a P2P Multicast Hybrid Overlay  
Tree Based on Topological Distances..... 307 - 321**  
*Sergej Alekseev and Jörg Schäfer*

# INFRASTRUCTURE CONSOLIDATION FOR INTERCONNECTED SERVICES IN A SMART CITY USING CLOUD ENVIRONMENT

Jorge F Hernandez<sup>1</sup>, Victor M Larios<sup>1</sup>, Manuel Avalos<sup>1</sup> and Ignacio Silva-Lepe<sup>2</sup>

<sup>1</sup>Department of Information Systems, CUCEA, UDG Guadalajara, Mexico  
jorge.hernandez.mx@ieee.org, victor.m.lariosrosillo@ieee.org,  
jmaavalos@mx1.ibm.com

<sup>2</sup>Thomas J. Watson Research Center, Yorktown Heights,  
NY USA Research, New York, USA  
isilval@us.ibm.com

## **ABSTRACT**

*Sustainability, appropriate use of natural resources and providing a better quality of life for citizens has become a prerequisite to change the traditional concept of a smart city. A smart city needs to use latest generation Information Technologies, IT, and hardware to improve services and data, to offer to create a balanced environment between the ecosystem and inhabitants. This paper analyses the advantages of using a private cloud architecture to share hardware and software resources when it is required. Our case study is Guadalajara, which has seven municipalities and each one monitor's air quality. Each municipality has a set of servers to process information independently and consists of information systems for the transmission and storage of data with other municipalities. We analysed the behaviour of the carbon footprint during the years 1999-2013 and we observed a pattern in each season. Thus our proposal requires municipalities to use a cloud-based solution that allows managing and consolidating infrastructure to minimize maintenance costs and electricity consumption to reduce carbon footprint generated by the city.*

## **KEYWORDS**

*Smart Cities; Cloud Architectures; Cost Estimation; City Services*

## **1. INTRODUCTION**

Improving the services offered by a city and promoting a balance between the environmental sustainability and citizen's quality of life has become an important goal of what we define today as Smart Cities [1]. IT offer a convenient way to connect processes in a city, optimize resources for the benefit of communities and forecast dynamics of the urban environment to better adapt solutions towards the well-being of citizens. However, citizens in smart cities have to deal with the physical and digital dimension.

During the living activities in the urban fabric, inhabitants have a unique identity to access and engage services such as energy, water, communication, and transport, among others. In addition, cities need to offer secure digital platforms for their inhabitants and IT infrastructure becomes vital in terms of communication and processing capabilities and availability. One solution to adapt and scale to the cities services demand is to shift city IT departments to the Cloud Computing paradigm [8].

The Cloud allows grouping various types of hardware and to merge them into a single entity for better and efficient management. Hence, Cloud Computing can work in three categories of services. First, we have the Infrastructure as a Service (IaaS) which provides of virtualization for using hardware resources and this category can offer sensors, storage or processing capacities on demand. Second is the Platform as a Service (PaaS) where users can run Web applications without the complexity of maintaining and running the associated infrastructure, this is critical for e-government service portals. Third, Software as a Service (SaaS) where licenses for critical software in processes such as analytics can be used on demand.

A key aspect of the cloud is the use of virtual machines to achieve its elasticity; a virtual machine is a software application that emulates be a real computer with software and hardware features limited to execute some task. Cloud computing provides a set of principles establishing the rules and principles among suppliers and customers. An important aspect of Cloud is the use of resources based on a “pay as you go” basis, in which the customer must pay for the time that a service, platform or software license is executed/used on a cloud provider.

A service as a process for a Smart City may need hardware, software or a combination of them. Cloud computing proposes benefits of elasticity, resilience, performance, productivity, scalability etc. Hence, this technology offers a better strategy for city governments to manage IT services. This paper is based on the Guadalajara Smart City project selected as IEEE pilot project to share the experience of best practices for smart cities worldwide. Moreover, Guadalajara is not only a city but also a metropolitan area composed of seven interconnected municipalities and we observed and analysed that each municipality has a traditional IT infrastructure consisting of a cluster of servers, routers and intranet access to communicate with other municipalities. In its current state, the data centers on each municipality are isolated infrastructure because they are not interconnected and sharing information and processing capabilities for the metropolitan area.

Our proposal is to consolidate municipal infrastructure by setting up a private cloud for the metropolitan area with the existing infrastructure. The benefits of using a cloud computing architecture allows the acquisition of any hardware configuration (supported by virtualization) in a few minutes. To better understand how Cloud Infrastructure can bring value for Smart Cities, we introduce a Use Case that is based on historical data about pollution in the metropolitan area; an alert system executed in the cloud can inform citizens when bad conditions can expose them to health threats. In Fig 1 we can see the core of the city which is deployed on a typical cloud computing architecture, in which a set of mobile devices or computers are interacting with users continuously to figure out how an alert system for city services can work on the cloud.



Fig. 1 A cloud infrastructure supporting City Services

## 2. PROBLEM STATEMENT

In the previous section, we discussed the benefits of Cloud Computing in its different layers (IaaS, PaaS, and SaaS) for Smart Cities. We also referred to the case of Guadalajara Smart City looking to shift from traditional IT infrastructure to a cloud computing environment to deal with the city dynamics.

We should mention that an additional important project is the interconnection of all government offices with optical fiber as per the project Connected Mexico, which offers to municipalities the conditions to share their IT infrastructure as a cloud entity. In Mexico, it is possible to process data outside the environment where they occur, i.e., each municipality can analyse data of citizens in another municipality; if and only if the citizen is informed how it will be used his personal data, thanks transparency and access to information law [11].

Current challenges in data centers include identifying the best practices to support a cloud-computing environment. Thus, we propose basic building blocks for this to migrate the traditional IT datacenters to a private cloud as shown in Fig 2. The hardware layer represents the physical resources (routers, computers, switches, hard drives, RAM memory, video cards, etc.) owned by the IT Municipal Datacenters. The second layer is the Virtualization, which enables create virtual machines when a process required it, with its own resources and its own operating system. The third layer has the software tools to complete integrate layers one and two.

For managing virtual machines can use applications such as OpenStack, VMware vSphere, CloudStack, Xen, among others. Government entities usually use open source solutions to minimize licensing costs nonetheless; they could use paid software to manage their virtual machines.

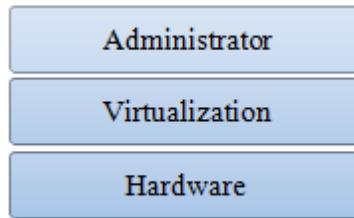


Fig. 2 Layers in a cloud computing environment

We can identify the Administrator as the process of monitoring the behaviour of each virtual machine in the cloud and to perform operations such as: increase or decrease resources hardware or software, delete, or create a new virtual machine. OpenStack Dashboard is an option for this type of module to support a better management. Given the layers in Fig. 2, Fig. 3 shows how the cloud service categories already explained fit into the Smart Cities cycles.

The city deploys sensors and actuators that can be connected to the cloud as an IaaS, offering a global management, security and capacity to scale on demand. In particular, sensors produce data to be curated and stored in the open data city repository. Cloud PaaS is the most indicated to curate and provide storage as well as processing capabilities for analytics. To deal with the complexity of Open Data repositories, specialized software for analytics should be used requiring a SaaS service to use licenses on demand.

One of the key elements of Smart Cities [4] is to break the silos of information among the different government offices offering to share all in a common open data city platform. This action allows avoiding duplicated efforts and investments to understand the city dynamics and provide solutions.

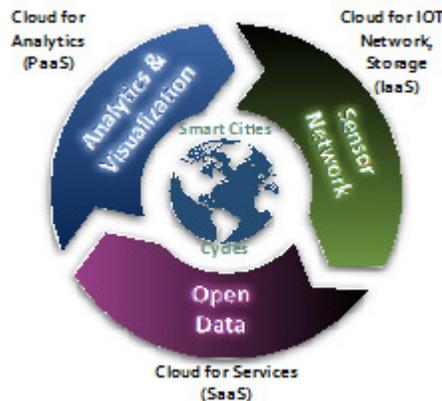


Fig. 3 Smart City Cycles related to the cloud

In order to provide more efficient services, it is possible to correlate information datasets from different indicators. Moreover, a Smart City needs to have a strategy for metrics to understand its performance and where to invest to reach its sustainability.

Metrics for smart cities need to have Key Performance Indicators (KPIs) as well as a subset of indicators for example in Cohen's Wheel there is a section called: Smart Gov, which contains infrastructure, this option could generate alternative metrics such as latency, multitenancy and others. This means that Smart Cities should work with a holistic vision integrating KPIs [5] and indicators to understand city dynamics and decide how the services should adapt. This concept is

based on a systemic approach where a city is a System of Systems or can be modelled and understood as a Complex System.

For this reason, the city should decide how to select KPIs and indicators. Given our work at the IEEE Smart Cities initiative, the model used is shown in Fig. 4 based in the Cohen Wheel. The model proposes five important KPIs related to Smart Economy, Smart Government, Smart People, Smart Living, Smart Mobility and Smart Environment [2][3]. Each KPI has a subset of actions and indicators in a secondary ring. It depends on the amount of sources of information available to feed indicators to be provided by the city, there could be more outer rings, which themselves induce more external rings.

This means that the more the city deploys sensor/actuator networks, the more rings that will appear, resulting in more accurate models to analyse the behaviours and dynamics of the city. That is the reason to have a good cloud strategy in order to scale the KPIs, Indicators and Actions management [6]. Hence, the Smart City project in Guadalajara, following the principles of Metrics based in the Cohen Wheel KPIs, requires an architecture to migrate the metropolitan area of Guadalajara to the cloud. This is the main problem and challenge presented in this paper.

A new issue to introduce is that the metropolitan area of Guadalajara, and for every city that is composed of interconnected municipalities, each one has autonomous infrastructure and budgets. Since all municipalities are interconnected, a challenge is to connect all data centers respecting their autonomy them. A proposed solution is to create a private cloud to support the three types of cloud services. As a use case to create a methodology to estimate the performance and cost of the private cloud integration among the interconnected municipalities, we identified sensors, open data and processing requirements as an example that can be used as reference for all KPIs of the Smart City in Guadalajara.

The sensors are real devices in the city creating datasets of air pollution in various zones of the city. The created datasets are stored in an open web service, and we propose a system that analyses the air pollution data flows to identify harmful pollution levels in zones of the metropolitan area to provide actions for the benefit of citizens (alarms, transport re-routing, etc.). The contribution of this paper is to process and analyse the information produced by an alert system. The system will be fully supported by the Cloud resulting in a consolidation of infrastructure across municipalities in the metropolitan area.

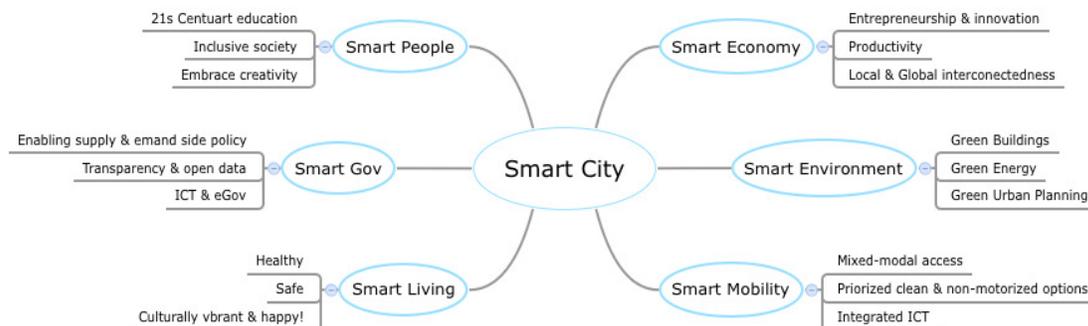


Fig. 4 Representative diagram Cohen's Smart City Wheel

Finally, we propose a methodology to estimate costs of cloud services, which is based in the current municipalities data center infrastructure modelled and extended with a plug in created for the Cloud Simulation Tool Cloudsim [7].

### 3. METHODOLOGY

We use Java framework to simulate the behaviour of a cloud, CloudSim. CloudSim's goal is to provide a generalized and extensible simulation framework that enables modelling, simulation, and experimentation of emerging Cloud computing infrastructures and application services, allowing its users to focus on specific system design issues that they want to investigate, without getting concerned with the low-level details related to Cloud-based infrastructure and services.

Also, an additional goal of our simulation was to generate an equation to create the actual cost of implementing the service using the cloud, we considered aspects such as payment of electricity, preventive and corrective maintenance, and key and support staff. We decided to group the different environments that are necessities to get the real cost of a specific service. These groups were:

1. *Physical configuration*: it represents the required configuration to execute correctly the service (i.e. hard disk, memory ram, video target, bandwidth, kind of network).
2. *Software configuration*: it refers to the set of programs that the service needs (i.e. operating system, database, file system, simulation programs and parallelism).
3. *Supplies*: items that the provider needs to active the mentioned above services (i.e. electrical power, cables, air conditioning, license fees, space, staff, payment to other providers).

This aggrupation allowed us to understand the elements to be evaluated in each process and we made an equation as follows:

$$C_{2ost}(\mu, \omega) = \sum_{i=1}^{\Psi} (\Omega_i + \phi_i) + \sum_{j=1}^{\mu} \lambda_{ij} + \sum_{k=1}^{\mathbf{Y}} \tau * C_{2ost}(\lambda, \tau) + M(\Omega)$$

where:

$\omega$  = a service/process

$\Psi$  = the total amount required resources

$\Omega$  = a specific used resource from physical or software configuration

$\phi$  = fixed cost of used resource finished by the provider

$\lambda$  = execution time of each resource

$\mu$  = execution time needed to complete all the process

$\mathbf{Y}$  = sub process of  $\omega$

$\mathbf{T}$  = the total sub process of  $\omega$

$M(\Omega)$  = maintenance cost of  $\theta$

Using this formula, we can obtain the cost computation of each resource in a specific time. We decided to use a recursive function to recover the used resources of a certain process/service that required a distribution of its job (*parallel tasks*). The primary goal of our formula is used to specify the economic cost in each process to have a log of all the physical resources (hardware) used in a service. Each municipality has its autonomy to decide what kind of software and hardware is needed to accomplish the tasks. With this equation we could also determine the efficiency of a process versus another for each municipality to identify the fastest execution and best low price service for the same service.

#### 4. EXPERIMENTAL FRAMEWORK

Guadalajara is located in western Mexico; it has a current population of about 4,299,000 according to the National Institute of Statistics and Geography (INEGI for its initials in Spanish) [9] in 2008. It is one of the three main cities in Mexico for its economic growth, technological and demographic region.



Fig. 5 Metropolitan area of Guadalajara<sup>1</sup>

The city is formed by 9 municipalities: Guadalajara, Zapopan, Tonalá, Tlaquepaque, El Salto, Juanacatlán, Ixtlahuacán de los Membrillos y Tlajomulco de Zuñiga. In Fig. 5, we show the structure the metropolitan area of Guadalajara. Since 1995, the city of Guadalajara has been monitoring air quality 24 hours a day, to make recommendations to care health of its citizens and animals. The Metropolitan Index of Air Quality [10] (IMECA in Spanish) is a Mexican official standard for gauging air quality since 1988. It reports chemicals such as: ozone (O<sub>3</sub>), sulfur dioxide (SO<sub>2</sub>), nitrogen dioxide (NO<sub>2</sub>), carbon monoxide (CO) and particulate less than 10 micrometers (PM<sub>10</sub>). IMECA is used as a reference for all the states of Mexico to measure environmental pollution. IMECA has a set of ranges for activating notifications, which are shown in the Table 1.

When the range of pollution is above 120 IMECAS, it may cause respiratory problems for some people, mainly children and the elderly. However, if the range is greater than 200 IMECAS a contingency plan is activated: the PRECA (Emergency Response Plan Contingency and

<sup>1</sup> Image taken from: [https://commons.wikimedia.org/wiki/File:Mapa\\_ZMG.svg](https://commons.wikimedia.org/wiki/File:Mapa_ZMG.svg)

Atmospheric Jalisco), which contains various levels, see Table 2. Guadalajara has registered as an open dataset, the information collected about pollution levels from 1996 to 2013, it is in a excel format and files contain data for each metropolitan area and its levels of pollution sensed per day and time (0 - 23 hours).

<i>IMECA</i>	<i>Air quality level</i>
0 – 50	Good
51 – 100	Regular
101 – 150	Bad
151 – 200	Very Bad
>201	Extremely Bad

Table 1. Classification pollution levels

<b>PRECA</b>		
<i>Level</i>	<i>Enabled</i>	<i>Disabled</i>
Pre	Equal or greater 120 IMECAS for 2 consecutive hours	Equal or less than 110 IMECAS for 2 consecutive hours
I	Equal or greater 150 IMECAS for 2 consecutive hours	Equal or less than 140 IMECAS for 2 consecutive hours
II	Equal or greater 200 IMECAS for 2 consecutive hours	Equal or less than 190 IMECAS for 2 consecutive hours
III	Equal or greater 250 IMECAS for 2 consecutive hours	Equal or less than 240 IMECAS for 2 consecutive hours

Table 2. Levels of activation and deactivation of PRECA

Registered contaminants are: O<sub>3</sub>, NO, NO<sub>2</sub>, NO<sub>x</sub>, SO<sub>2</sub>, CO, PM<sub>10</sub>, direction and wind speed and temperature. In our case study, we considered the use of CO as a contaminant, because most Guadalajara inhabitants are in direct contact with it. The CO is an odourless and invisible toxin, thus constant attention is required to monitor its level and to detect when it is higher than environmental health standards in Mexico, to activate contingency plans for the population. The extracted information on the levels of contamination of the ZMG, which is in Excel format, it was migrated to a MySQL database, we normalized the information and we analysed the behaviour of the contamination by zone based on hours, days, months and seasons.

In Figs. 6 to 8, we can observe the behaviour of contamination of three major areas of Guadalajara, where we no increase in pollution levels at the beginning of January, May and December. A decrease in the months of August and September. These factors should be for the rainy season and the reincorporation of students and teachers to their schools. We noted that there is a relationship between the work and school activities with the emission of pollutants, with this premise determined in 3 phases: 1 = low level, 2 = intermediate level and 3 = maximum level. We mapped these levels with an alert system (we did it) to inform to the citizens and to initiate a feedback process to provide recommendations and avoid crossing by this zone.



Fig. 6 Co Levels in Las Pintas 2013

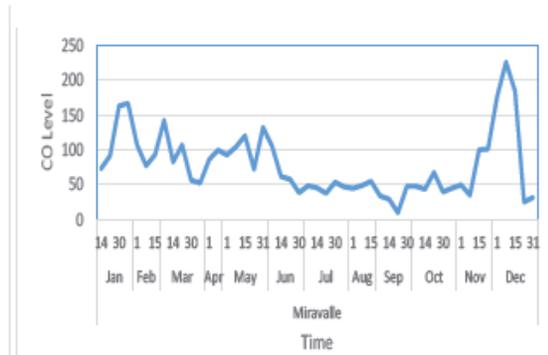


Fig. 7 Co Levels in Miravalle 2013



Fig. 8 Co Levels in Tlaquepaque 2013

These levels allow us to identify when a municipality zone needs to share resources among others through the private cloud to manage services. When a level is in phase 1, the municipality may process its information by itself, in phase 2, it could need the assistance of one or more municipalities to improve response time, and the last phase this municipality requests the intervention of others to process the high demand for information to be processed in the shortest possible time. Once the simulation has been performed, we create nine datacenters with hardware features provided by each zone also. The purpose of representing the nine municipalities in a cloud environment, it is to verify that sharing the resources of each infrastructure, improving quality of service to the city, maintenance costs are optimize and the purchase of computer equipment benefits all municipalities and not just one. That is reason, we generated a mechanism to share resources between each area and we built a coordinator agent, which monitors the workload of each area to grant access to other data centers when is necessary.

After we configured a datacenter with the following characteristics: 16 Memory Ram, 4 Tb Hard Drive, Two Xeon X3430 processors and multi-node optical fiber with 100 Mbps transfer rate. This configuration represents the average of resources that a municipality has. The application was executed in two phases: the first, it was using the traditional scheme of each municipality and phase two: it used a cloud-based scheme. In the Fig. 9 we show the workflow, we did for this extraction, interpretation and results obtained.

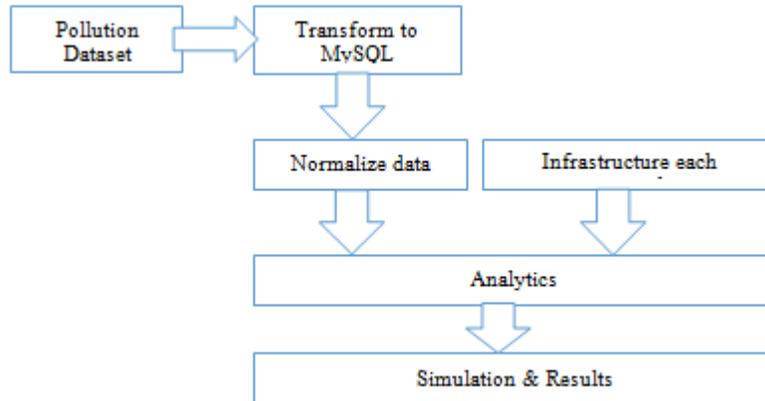


Fig. 9 Workflow process

### 5. RESULT ANALYSIS

In Fig 10, we show on the left side the number of milliseconds that a service took to execute during a certain time in a traditional scheme, and on the right side the result of our proposal where we suggested use of resources using a private cloud. A considerable decrease is observed and thus electric power consumption was reduced and the cost of use is lower than with other scheme.

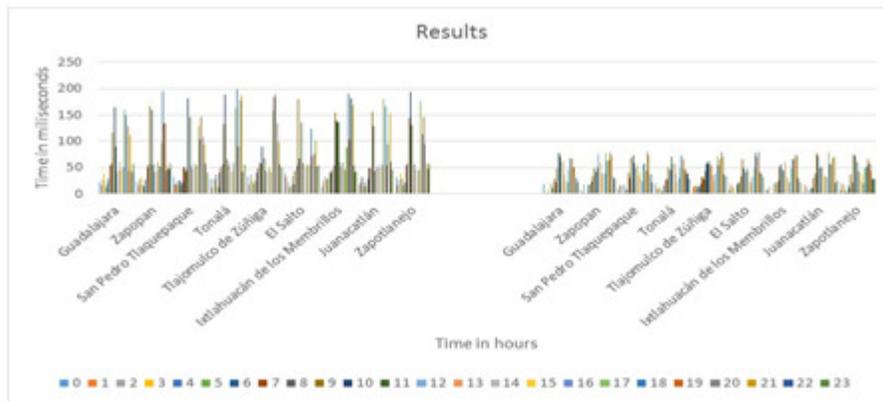


Fig. 10 Traditional IT vs Private Cloud

## 6. CONCLUDING REMARKS AND PERSPECTIVES

Historical data of pollution levels helped to determine the system of events that should be present in the city to inform and organize services for each zone. For example, when pollution levels are high, systems could report to citizenship through their smartphones of this situation and suggest changes in their routine. Thus begins a process of feedback where the user could request new routes to arrive his destination not passing by contaminated areas.

These kind of services are useful for smart cities. Cloud computing allows enhancing resources and adapt to new hardware equipment that may be present in a data center and to share its resources with other areas. As a future work, we could use docker containers to perform in situ the information that is enclosed within a metropolitan area instead of moving it to another location for processing.

## ACKNOWLEDGEMENTS

The work described in this paper was supported by CONACYT through University of Guadalajara in collaboration with Smart Cities Innovation Center.

## REFERENCES

- [1] M. Batty, K. W. Axhausen, F. Giannotti, A. Pozdnoukhov, A. Bazzani, M. Wachowicz, G. Ouzounis, and Y. Portugali, "Smart cities of the future," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 481–518, Nov. 2012.
- [2] H. Schaffers, N. Komninos, M. Pallot, B. Trousse, M. Nilsson, and A. Oliveira, "Smart Cities and the Future Internet: Towards Cooperation Frameworks for Open Innovation.," *Future Internet Assembly*, vol. 6656, no. 31, pp. 431–446, 2011.
- [3] "Smart city Framework. Guide to establishing strategies for smart cities and communities," BSI British Standards, London.
- [4] N. Komninos, H. Schaffers, and M. Pallot, "Developing a Policy road map for Smart Cities and the future internet," *eChallenges e-2011 ...*, 2011.
- [5] C. Harrison and I. A. Donnelly, "A Theory of Smart Cities," *Proceedings of the 55th Annual Meeting of the ISSS - 2011*, Hull, UK, vol. 55, no. 1, Sep. 2011.
- [6] G. J. Peek and P. Troxler, "City in Transition: Urban Open Innovation Environments as a Radical Innovation," *programm.corp.at*
- [7] Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, Cesar A. F. De Rose, and Rajkumar Buyya, "CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms, Software: Practice and Experience (SPE)", Volume 41, Number 1, Pages: 23-50, ISSN: 0038-0644, Wiley Press, New York, USA, January, 2011
- [8] Rajkumar Buyya, James Broberg, Andrzej M. Goscinski , "Cloud Computing: Principles and Paradigms", Wiley Editorial, March 2011
- [9] Instituto Nacional de Estadística Geografía e Informática, <http://www.inegi.org.mx/>

[10] Secretaria de Medio Ambiente, <http://siga.jalisco.gob.mx/>

[11] Ley Federal de Transparencia y Acceso a la Información Pública, [http://www.diputados.gob.mx/LeyesBiblio/pdf/244\\_140714.pdf](http://www.diputados.gob.mx/LeyesBiblio/pdf/244_140714.pdf), Junio 2002

## AUTHORS

**Jorge F. Hernandez**, he was born in Guadalajara, Jalisco, in 1975. He received the B.E. degree in computation engineering from the University of Guadalajara, Mexico, in 1998, and the Master in Computer Science in 2000 from Cinvestav. In 2004, he joined the Department of Computer Science, University of Guadalajara, as a teacher. He had published 4 articles in different congress and he had been thesis advisor in bachelor and master level. In 2012, he was admitted in CUCEA to study Postgrade IT and he is working with cloud computing to estimate cost and planning time processing.



**Víctor M. Larios**, has received his PhD and a DEA (French version of a MS program) in Computer Science at the Technological University of Compiègne, France and a BA in Electronics Engineering at the ITESO University in Guadalajara, Mexico. He works at the University of Guadalajara (UDG) as Professor Researcher and as consultant directed the Guadalajara Ciudad Creativa Digital Science and Technology program during 2013. His research interests are related to distributed systems, visual simulations and smart cities. He is a Senior IEEE member and current chair of the Computer Chapter at the IEEE Guadalajara Section at Region 9. His role in the IEEE-CCD Smart Cities initiative is to lead the working groups.



**Manuel Avalos**, graduated from Computer Science Engineering from “Universidad Autónoma de Guadalajara (UAG)” back in 1991. In December 2006 Manuel finished his Master Degree on Information Technology by ITESM institution. Manuel is currently a PhD Data Science student at Universidad de Guadalajara campus CUCEA. Manuel joined IBM in 1991 as a Testing Software Engineer and since then Manuel had several Technical, Management and Executive positions in different IBM Divisions, currently he has a Global responsibility for Systems-Storage brand.



**Ignacio Silva-Lepe** is a Research Staff Member at IBM. His areas of interest include (1) Component Software, designing and building application server infrastructure for distributed components, (2) Distributed Messaging Systems, (3) Advanced Enterprise Middleware, and (4) PaaS Research, designing and building infrastructure for on-boarding and instantiating platform as a service offering onto a compute cloud. Before joining IBM Research, Ignacio was a Member of Technical Staff at Texas Instruments' Corporate Research and Development, subsequently acquired by Raytheon. Prior to that he was a Research Assistant at Northeastern University, where he earned a PhD in Computer Science.



# SIMILARITY ANALYSIS OF DNA SEQUENCES BASED ON THE CHEMICAL PROPERTIES OF NUCLEOTIDE BASES, FREQUENCY AND POSITION OF GROUP MUTATIONS

Fatima KABLI<sup>1</sup>, Reda Mohamed HAMOU<sup>2</sup>, Abdelmalek AMINE<sup>3</sup>

GeCode Laboratory, Department of Computer Science  
Tahar MOULAY University of Saïda, Algeria.

<sup>1</sup>kablifatima47@gmail.com, <sup>2</sup>hamoureda@yahoo.fr,  
<sup>3</sup>amine\_abdl@yahoo.fr

## **ABSTRACT**

*The DNA sequences similarity analysis approaches have been based on the representation and the frequency of sequences components; however, the position inside sequence is important information for the sequence data. Whereas, insufficient information in sequences representations is important reason that causes poor similarity results. Based on three classifications of the DNA bases according to their chemical properties, the frequencies and average positions of group mutations have been grouped into two twelve-components vectors, the Euclidean distances among introduced vectors applied to compare the coding sequences of the first exon of beta globin gene of 11 species.*

## **KEYWORDS**

*DNA sequence, chemical proprieties, DNA bases, position, frequency, group mutations.*

## **1. INTRODUCTION**

DNA Sequence similarity is fundamental challenge in bioinformatics to predicting unknown sequences functions or effects, constructing phylogenetic tree, and identify homologous sequences, several of DNA sequence similarity measuring approaches have been developed, divided into several categories, the alignment-based, alignment-free, statistics method and others.

Most methods based on the concept of the sequence alignment defined as a way of arranging the sequences of DNA, RNA, or proteins to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences, such as BLAST[1] (Basic Local Alignment Search Tool), was developed as a way to perform DNA and protein sequence similarity searches, is a heuristic method considered as a rapid approach for sequence comparison, based on the comparaison of all combinations of nucleotide or protein queries with nucleotide or protein databases. there are several types of BLAST, Another method call Fasta [2]

uses the principle of finding the similarity between the two sequences statistically. This method matches one sequence of DNA or protein with the other by local sequence alignment method. It searches for local region for similarity and not the best match between two sequences.

In addition to performing alignments, is very popular due to its availability on the World Wide Web through a large server at the National Center for Biotechnology Information (NCBI) and at many other sites. Has evolved to provide molecular biologists with a set of very powerful search tools that are freely available to run on many computer platforms.

Also, they are UCLUST [3] and CD-HIT [4] and many more, Obviously it consumes time while running however, the similarity can be quickly computed with the alignment-based method that converts each piece of DNA sequence into a feature vector in a new space. To generate feature vectors some algorithms exploit probabilistic models of which the Markov model [5-6], SVM-based approaches [7], widely used in bioinformatics applications.

Other technique used statistics method for sequence comparison, based on the joint k-tuple content in two sequences called K-tuple Algorithm One of the very popular alignment-free methods [8, 9], in which DNA sequence is divided into a window of length k (word of length). The feature vector is generated by the calculated to the frequency value of each tuple; the similarity can be quickly measured by some distance metric between vectors. Such as KLD [10] from two given DNA sequences, was constructed two frequencies vectors of n-words over a sliding window, whereas was derived by a probabilistic distance between two Sequences using a symmetrized version of the KLD, Which directly compares two Markov models built for the two corresponding biological sequences.

On the other hand, these methods cannot completely describe all information contained in a DNA sequence, since they only contains the word frequency information, therefore, many researches modified k-tuple are proposed to contain more information. [11] used both the overlapping structure of words and word frequency to improve the efficiency of sequence comparison. [12] Transformed the DNA sequence into the 60-dimension distribution vectors.

In order to help improve DNA sequence analysis methods ,the graphical representations of DNA sequences on 2D or 3D space [13-14] applied by several researches, but there are some disadvantage as loss of information due to crossing and overlapping of the curve representing DNA with itself [15-16]. To avoid this problem many new graphical representation methods recently [17-14] have been invented.

Other works [18-19] have based on the dinucleotide analysis. To reveal the biology information of DNA sequences. Based on qualitative comparisons used the three classifications of the four DNA bases A, G, T and C, according their chemical properties.

[20] Present the DNA sequence by a 12-component vector consisting of twelve frequencies of group mutations, and calculated the similarity between deferent vectors by the Euclidian distance. While [21], converted a DNA sequence into three 2-dimension cumulative ratio curves the R/Y-ratio curve, the K/M-ratio curve and the W/S-ratio curve, the coordination of every node on these 2-D cumulative ratio curves have clear biological implication.

Li and Wang [22] presented a 16-dimension binary vector based also on the group of nucleotide bases. These methods give encouraging results, they are focused much more on the sequence frequency than the position for sequences analyses, Dong and Pei [23] argued that the position inside sequence is important information. Therefore, insufficient information in a feature vector is an important reason that causes poor similarity results.

In this paper, we combine the advantages of other methods with our own proposal. We presented each DNA sequence by three symbolic sequences according to their chemical properties, the group mutations have been grouped into two twelve-components vectors. The first represents the frequency and the second represents the average position, to compare the coding sequences of the first exon of beta globin gene of 11 different species, we applied the Euclidean distances among introduced vectors.

## 2. MODELLING

### 2.1 Data Set

We have used in our experimentation DNA sequences derived by the data Set obtained by [23], the data set contains the first exon of beta globin genes of 11 different species in Table 1

Table 1. The first exon of beta globin genes of 11 different species

Species	Sequences
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTACC GGCTTCTGGGGCAAGGTGA AAGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Opossum	ATGGTGCACCTGACTTCTGAGGAGAAGAAGTGCATCACTACCATCTGGTCTAA GGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTGGGCAG
Gallus	ATGGTGCACCTGACTGCTGAGGAGAAGCAGCTCATCACC GGCTTCTGGGGCA AGGTCAATGTGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTACCTCTCTGTGGGGCAA GGTGGATGTAGAGAAAGTTGGTGGCGAGGCCCTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGGCAA AGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGACGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAA GGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTACC GGCTTTTGGGGCAAGGTGAA AGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGG

## 2.2 Proposed Method

It is difficult to obtain the information from DNA primary sequence directly; in our approach we based on the three classes of DNA bases, according the chemical properties, the purine group  $R = \{A, G\}$  and pyrimidine group  $Y = \{C, T\}$ ; amino group  $M = \{A, C\}$  and keto group  $K = \{G, T\}$ ; weak H-bond group  $W = \{A, T\}$  and strong H-bond group  $S = \{C, G\}$ . They call RY classification, MK classification and WS classification correspondingly. Whereas, for the primary sequence  $X = S_1..S_2...S_3.... S_n$ , with length  $n$ , is presented by three different sequences according the three classification, RY, MK and WS by  $\Phi_{RY}, \Phi_{MK}, \Phi_{WS}$

$$\Phi_{RY}(X) = \Phi_{RY}(S_1) \Phi_{RY}(S_2) \dots \Phi_{RY}(S_n), \Phi_{MK}(X) = \Phi_{MK}(S_1) \Phi_{MK}(S_2) \dots \Phi_{MK}(S_n), \Phi_{WS}(X) = \Phi_{WS}(S_1) \Phi_{WS}(S_2) \dots \Phi_{WS}(S_n).$$

$$\text{Where } \Phi_{RY}(S_i) = \begin{cases} R & \text{if } S_i \in R \\ Y & \text{if } S_i \in Y \end{cases}, i=1,2,\dots,n \quad (1)$$

$$\text{Where } \Phi_{MK}(S_i) = \begin{cases} M & \text{if } S_i \in M \\ K & \text{if } S_i \in K \end{cases}, i=1,2,\dots,n \quad (2)$$

$$\text{Where } \Phi_{WS}(S_i) = \begin{cases} W & \text{if } S_i \in W \\ S & \text{if } S_i \in S \end{cases}, i=1,2,\dots,n \quad (3)$$

Each DNA sequence represented by the three symbolic sequences according the three formula above.

### 2.2.1 Frequency Analyses

In each classification we focus on group mutation information, for the three symbolic sequences there are twelve group mutations,  $R \rightarrow R, R \rightarrow Y, Y \rightarrow R, Y \rightarrow Y, M \rightarrow M, M \rightarrow K, K \rightarrow M, K \rightarrow K, W \rightarrow W, S \rightarrow W, W \rightarrow S, S \rightarrow S$ .

As a first step, we calculated the frequency of each mutation information defined by the following formula used by [19].

$$f_{UV} = \frac{\text{the number of word UV}}{n-1} \quad (4)$$

UV is the mutation information for RY classification, the frequencies denoted by  $f_{RY} f_{RY} f_{YR} f_{YY}$ , while the frequencies of MK classification denoted by  $f_{MM} f_{MK} f_{KM} f_{KK}$ , and for WS classification denoted by  $f_{WW} f_{WS} f_{SW} f_{SS}$ .

The table 2 present the frequencies of group mutations of the first exon of  $\beta$ -globin gene of eleven species based on the three symbolic sequences of DNA.

Table 2. Frequencies of group mutations of 11 species

	$f_{RR}$	$f_{RY}$	$f_{YR}$	$f_{YY}$	$f_{MM}$	$f_{MK}$	$f_{KM}$	$f_{KK}$	$f_{WW}$	$f_{WS}$	$f_{SW}$	$f_{SS}$
Human	0.3297	0.2308	0.2308	0.2088	0.1978	0.1978	0.1868	0.4176	0.1209	0.2967	0.2857	0.2967
Gallus	0.3407	0.2308	0.2308	0.1978	0.2418	0.2308	0.2198	0.3077	0.0989	0.2747	0.2637	0.3626
Lemur	0.3626	0.2198	0.2198	0.1978	0.1319	0.2418	0.2308	0.3956	0.1429	0.3187	0.3077	0.2308
Rabbit	0.3483	0.2472	0.2360	0.1685	0.1798	0.1910	0.1910	0.4382	0.1011	0.3146	0.3034	0.2809
Rat	0.3297	0.2418	0.2418	0.1868	0.1978	0.2198	0.2088	0.3736	0.1758	0.2747	0.2637	0.2857
Bovine	0.3882	0.2118	0.2118	0.1882	0.1765	0.2118	0.2000	0.4118	0.1294	0.2824	0.2706	0.3176
Opossum	0.3077	0.2308	0.2308	0.2308	0.2308	0.2198	0.2088	0.3407	0.1319	0.3407	0.3297	0.1978
Gorilla	0.3478	0.2283	0.2283	0.1957	0.1957	0.1957	0.1848	0.4239	0.0978	0.3043	0.2935	0.3043
Mouse	0.3118	0.2258	0.2258	0.2366	0.1935	0.2043	0.1935	0.4086	0.1183	0.3118	0.3011	0.2688
Goat	0.3882	0.2118	0.2118	0.1882	0.1647	0.2353	0.2235	0.3765	0.1059	0.2941	0.2824	0.3176
Chimpanzee	0.3462	0.2308	0.2308	0.1923	0.1923	0.1923	0.1827	0.4327	0.1250	0.2981	0.2885	0.2885

### 2.2.2 Position analyses

The position inside sequence is important information Therefore; insufficient information in a feature vector is important reason that causes poor similarity results.

For instance, if two sequences have the same frequency of components but have two different sequencing directions, if we calculate just the frequency similarity. We get them identical, but the position of their components is completely different, there is no biological relationship between them. For this reason and to improve the effectiveness of similarity study of DNA sequence, that considered as a main challenge in the field of bioinformatics sequences. We used the concept of the position of the DNA components.

We based on the group mutations presented above for calculate the average position (average distance); we have proposed the following formula.

$$P_{UV} = \frac{(\sum_{i=1}^k (Position_{UV}))}{K*(n-1)} \quad (5)$$

K is the number of word UV.

Wherein the position of each component is defined as the average position of the word uv divide by the length of the DNA sequence n.

The Table 3 present the average position of group mutations for the first exon of  $\beta$ -globin gene of eleven species based on the three symbolic sequences of DNA.

Table 3. Average Position of group mutations of 11 species.

	$P_{RR}$	$P_{RY}$	$P_{YR}$	$P_{YY}$	$P_{MM}$	$P_{MK}$	$P_{KM}$	$P_{KK}$	$P_{WW}$	$P_{WS}$	$P_{SW}$	$P_{SS}$
Human	0.5502	0.4746	0.4956	0.4274	0.4621	0.4512	0.4551	0.5480	0.5325	0.4554	0.4573	0.5539
Gallus	0.5115	0.4558	0.4762	0.5317	0.5490	0.4668	0.4670	0.4922	0.5336	0.4259	0.4286	0.5837
Lemur	0.5608	0.4632	0.4841	0.4194	0.5220	0.4590	0.4636	0.5250	0.4632	0.4505	0.4505	0.6332
Rabbit	0.5814	0.4479	0.4414	0.4569	0.4789	0.4243	0.4613	0.5457	0.4969	0.4539	0.4557	0.5807
Rat	0.5326	0.4500	0.4695	0.5171	0.4847	0.4879	0.4922	0.5048	0.4457	0.4585	0.4592	0.5917
Bovine	0.5387	0.4654	0.4876	0.4419	0.5192	0.4163	0.4187	0.5600	0.4920	0.4716	0.4747	0.5316
Opossum	0.5165	0.4731	0.4950	0.4861	0.4987	0.4599	0.4610	0.5346	0.3974	0.4747	0.4751	0.6258
Gorilla	0.5635	0.4695	0.4896	0.4070	0.4571	0.4463	0.4501	0.5535	0.4855	0.4596	0.4622	0.5637
Mouse	0.5877	0.4424	0.4644	0.4506	0.5269	0.4493	0.4528	0.5218	0.4399	0.4520	0.4531	0.6146
Goat	0.5258	0.4654	0.4876	0.4684	0.5277	0.4382	0.4409	0.5460	0.5033	0.4701	0.4735	0.5316
Chimpanzee	0.5502	0.4780	0.4956	0.4163	0.4606	0.4514	0.4555	0.5468	0.5851	0.4587	0.4603	0.5288

### 2.2.3 Example

For the Following Sequence

ATGGTGACCTGAC

We get the three symbolic sequences:

$$\Phi_{RY} = \mathbf{RYRRYRYRYYYRRY.}$$

$$\Phi_{MK} = \mathbf{MKKKKKMMMMKKMM.}$$

$$\Phi_{WS} = \mathbf{WWSSWSSWSSWSWS.}$$

From the three sequences we constructed two twelve-component vectors, the first for calculate the frequency of group mutations and the second for their average position.

For the Word RR, its frequency calculated by the formula (4)  $F_{(RR)} = \frac{2}{(14-1)} = 0.15$ , and its average position based on formula (5) is  $P_{(RR)} = \frac{((3+12)/2)}{(14-1)} = 0.57$ , the same thing for the rest group mutations,

### 2.3 Similarity and Dissimilarity

In order to analysis the similarity and dissimilarity between two DNA sequences, each sequence represented by two twelve-component vectors as presented above, The similarities between such vectors calculated by the Euclidian distance between their end points for both vectors frequency and average position. In the next, we calculated the average distance by the following formula.



Table 6. Average Similarity/dissimilarity matrix between frequency and position of group mutations for the 11 genes sequences.

	Human	Gallus	Lemur	Rabbit	Rat	Bovine	Opossum	Gorilla	Mouse	Goat	Chimpanzee
Human	0.0	0.1563	0.1260	0.0791	0.1209	0.0877	0.1624	0.0457	0.0980	0.099	0.0451
Gallus		0.0	0.1847	0.1705	0.1316	0.1578	0.1902	0.1718	0.1644	0.1293	0.1771
Lemur			0.0	0.1136	0.1270	0.1295	0.1278	0.1186	0.0873	0.1229	0.1484
Rabbit				0.0	0.1253	0.1013	0.1601	0.0672	0.0917	0.1112	0.0909
Rat					0.0	0.1338	0.1242	0.1318	0.1110	0.1266	0.1485
Bovine						0.0	0.1772	0.0842	0.1242	0.0524	0.1001
Opossum							0.0	0.1583	0.1134	0.1677	0.1957
Gorilla								0.0	0.0962	0.1017	0.0711
Mouse									0.0	0.1287	0.1326
Goat										0.0	0.1150
Chimpanzee											0.0

We have observed in (Table 6) of similarity above, that is a great similarity between the sequence of human with gorilla, human with Chimpanzee and chimpanzee with gorilla another similarity between mouse, lemur, and mouse with rat, such as mouse and rat belong to the same Muridae mammalian family. Also for the bovine and goat, they belong to the same Bovidae mammalian family.

Each of opossum and Gallus are far from the rest species, because opossum is the most remote species from the remaining mammals and the Gallus is the only non-mammalian animal among all other species of the dataset. However, the rest nine species are mammals family.

The obtained result it is not an accident, but shows the relationship in evolutionary sense between the twelve species.

The relationship between the 11 species according our own DNA analysis presented in the following dendrogram.

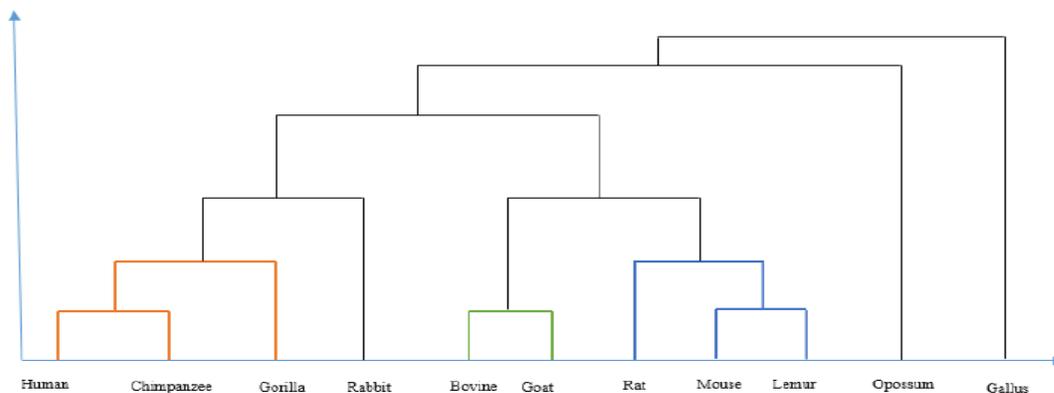


Figure 1. The dendrogram of twelve species.

### 3. CONCLUSION AND PERSPECTIVE

Similarity analysis of DNA sequences are still important subjects in bioinformatics, the similarity between two DNA sequences defined by the frequency and position of their components. The representation of a DNA sequence by three symbolic sequences helpful to define all possible mutations groups. We build two-dimensional vectors, the first represents the frequency of mutation groups and the second represents their average positions,

To calculate the similarity and dissimilarity of DNA sequences, Euclidean distances are applied based on the frequency and position of mutation groups

The evaluation results of 11 different species coincides with the evolutionary sense. The proposed method has a wide Range of applicability for analysis of biological sequence.

### REFERENCES

- [1] Gish W, Miller W, Myers E, Lipman D, AltschulS: Basic local alignment search tool. *J Mol Biol* , 215(3):403-410. doi:10.1016/S0022-2836(05)80360-2 (1990).
- [2] Lipman DJ, Pearson WR: Rapid and sensitive protein similarity searches. *Science*, 227:1435-1441, (1985).
- [3] Edgar RC: Search and clustering orders of magnitude faster than blast *Bioinformatics*, 26:2460-2461, (2010).
- [4] Li WZ, Godzik A: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658-1659, (2006).
- [5] Pham TD, Zuegg J: A probabilistic measure for alignment-free sequence comparison. *Bioinformatics* , 20:3455-3461, (2004).
- [6] Freno A: Selecting features by learning markov blankets. *Lect Notes Comput Sci*, 4692:69-76, (2007).
- [7] Deshpande M, Karypis G: Evaluation of techniques for classifying biological sequences. *Lect Notes Comput Sci*, 2336:417-431, (2002).
- [8] Blaisdell BE: A measure of the similarity of sets of sequences not requiring sequence alignment. *Proc Natl Acad Sci U S A*, 83(14):5155-5159, (1986).
- [9] Vinga S, Almeida J: Alignment-free sequence comparison—a review. *Bioinformatics*, 19:513-523, (2003).
- [10] Wu,T.J.,Hsieh,Y.C.and Li,L. A. Statistical measures of DNA dissimilarity under Markov chain models of base composition. *Biometrics*, 57,441–448, (2001).
- [11] Dai Q, Liu XQ, Yao YH, Zhao FK: Numerical characteristics of word frequencies and their application to dissimilarity measure for sequence comparison. *J Theor Biol* , 276:174-180, (2011).
- [12] Zhao B, He RL, Yau SS: A new distribution vector and its application in genome clustering. *Mol Phylogenet Evol* , 59:438-443, (2011).
- [13] Hamori, E., Ruskin, J., Curves, H.: A Novel Method of Representation of Nucleotide Series Especially Suited for Long DNA Sequences. *J. Biol. Chem.* 258, 1318–1327 (1983).
- [14] Qi, Z., Qi, X.: Novel 2D graphical representation of DNA sequence based on dual nucleotides. *Chem. Phys. Lett.* 440, 139–144 (2007).
- [15] Gates, M.A.: A Simple way to look at DNA. *J. Theor. Biol.* 119, 319–328 (1986).
- [16] Guo, X.F., Randic, M., Basak, S.C.: A novel 2-D graphical representation of DNA sequences of low degeneracy. *Chem. Phys. Lett.* 350, 106–112 (2001).
- [17] Randic, M., Vrakoc, M., Lers, N., Plsvsic, D.: Novel 2-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* 368, 1–6 (2003).
- [18] Liu, X.Q., Dai, Q., Xiu, Z.L., Wang, T.M.: PNN-curve: A new 2D graphical representation of DNA sequences and its application. *J. Theor. Biol.* 243, 555–561 (2006).

- [19] i, Z., Fan, T.: PN-curve: A 3D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.* 442, 434–440 (2007).
- [20] Shi L, Huang HL: Dna sequences analysis based on classifications of nucleotide bases. *Adv Int Soft Comput* , 137:379-384, (2012).
- [21] Yu HJ: Similarity analysis of dna sequences based on three 2-d cumulative ratio curves. *Lect Notes Comput Sci* , 6840:462-469, (2012).
- [22] Li C, Wang J: Similarity analysis of dna sequences based on the generalized lz complexity of (0,1)-sequences. *J Math Chem* , 43:26-31, (2008).
- [23] Dong GZ, Pei J: Classification, clustering, features and distances of sequence data. *Adv Database Syst*, 33:47-65, (2007).
- [24] Nandy, A., Harle, M., Basak, S.C.: Mathematical descriptors of DNA sequences development and applications. *ARKIVOC* ix, 211–238 (2006).
- [25] Zhao, L., et al.: An S-Curve-Based Approach of Identifying Biological Sequences. *Acta Biotheoretica* 58(1), 1–14 (2009).
- [26] Xie, G., Mo, Z.: Three 3D graphical representations of DNA primary sequences based on the classifications of DNA bases and their applications. *Journal of Theoretical Biology* 269(1), 123–130 (2011).
- [27] Wu TJ, Huang YH, and Li LA, Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences. Vol. 21 no .222005, pages 4125–4132 doi: 10.1093/bioinformatics/bti658, (2005).
- [28] Sierk M, Person W. Sensitivity and Selectivity in Protein Structure Comparison. *Protein Sci.*2004 ; 13:773–785.
- [29] Krasnogor N, Pelta DA. Measuring the Similarity of Protein Structures by Means of the Universal Similarity Metric. *Bioinformatics.* 2004;20:1015–1021.
- [30] Reinert G, Schbath S, and Waterman MS. Probabilistic and statistical properties of words: an overview. *J Comput Biol.* 2000;7:1–46.
- [31] QI. D and Wang T , Comparison study on k-word statistical measures for protein: From sequence to 'sequence space'

# ABOUT THE SUITABILITY OF CLOUDS IN HIGH-PERFORMANCE COMPUTING

Harald Richter

Clausthal University of Technology, Clausthal, Germany  
hri@tu-clausthal.de

## **ABSTRACT**

*Cloud computing has become the ubiquitous computing and storage paradigm. It is also attractive for scientists, because they do not have to care any more for their own IT infrastructure, but can outsource it to a Cloud Service Provider of their choice. However, for the case of High-Performance Computing (HPC) in a cloud, as it is needed in simulations or for Big Data analysis, things are getting more intricate, because HPC codes must stay highly efficient, even when executed by many virtual cores (vCPUs). Older clouds or new standard clouds can fulfil this only under special precautions, which are given in this article. The results can be extrapolated to other cloud OSES than OpenStack and to other codes than OpenFOAM, which were used as examples.*

## **1. INTRODUCTION**

Cloud computing has become the new ubiquitous computing and storage paradigm. Reasons are e.g. the pay-as-you-go accounting, the inherent elasticity, flexibility and user customizing. An example therefore is the arbitrary creation and deletion of virtual machines (VMs) with individual numbers of vCPUs and flexible amounts of virtual main memory per VM. Companies, institutions and individuals can profit from this paradigm by off-loading computing and storage to commercial cloud service providers (CSPs), because CSP services range from simple data backups to entire virtual data centres. These advantages make cloud computing also attractive for scientists, since they do not need any more to provide and maintain their own computer infrastructure, but can out-source it to CSPs, which is then hosted as virtual IT. However, for the case of High-Performance Computing (HPC) in a cloud, as it is needed in simulations or Big Data analysis, things are getting more intricate, because HPC codes must stay highly efficient and thus scalable, even when executed by many virtual cores (vCPUs), which are located on different physical servers. This is not necessarily the case in older clouds and also not in newer standard clouds, as measurements made by several research groups have shown. For example, the US Dep. of Energy (DoE), which is responsible for HPC in the USA, has questioned the usefulness of clouds for HPC in 2011 [1] in general. Other authors with the same opinion are [2]-[4], for example. We believe that further research effort is needed to improve the execution efficiency and the speed-up for HPC in clouds, but it is also our opinion that this is possible. In this contribution, multiple reasons for cloud inefficiencies are presented for the case of OpenStack [5] as cloud operating system and for OpenFOAM [6] as HPC example code. Suggestions are made how to solve or circumvent them. The results can be extrapolated to other cloud OSES and other HPC codes. The contribution is organized as follows: In chapter 2, the state-of-the-art is reviewed. Chapter 3 describes our project and what equipment and tools we were using. In

chapter 4, the conducted measurements and findings are presented and discussed. The paper ends with a conclusion, followed by an outlook and a reference list.

## 2. STATE-OF-THE-ART

Several HPC-related scientific projects dedicated to cloud optimization and scheduling were studied by us, and a thorough literature search was conducted to determine which characteristics of the cloud are needed to make it HPC-capable. The papers, that are reviewed here are [3], [4] and [7]-[11] (in rank of importance as we felt it). Also the author of this paper has worked previously on cloud performance issues in [12] and in [14]. In [3], [4] and in this article, it is stated that the overhead in virtualized communication can lead to bad performance because of inter-server bottlenecks. This is true if not the latest hardware-accelerators for virtualized communication are used. Furthermore, according to the studied literature, existing cloud schedulers ignore the needs HPC codes have, as well as the heterogeneity and the multi-tenancy clouds have with respect to their resources. It is furthermore said that these are the major cloud-intrinsic bottlenecks that prevent from effective HPC.

**Scheduling:** In order to address the HPC bottlenecks, the authors of [3], [4], for example, have enhanced the Nova scheduler of OpenStack with the information that a job is a HPC application. Together with other meta-information for Nova, they achieved a performance improvement of 45%, which is a remarkable result. In [7], a monograph on scheduling approaches is given, which is also relevant for clouds. In [8], inter-cloud meta-schedulers are discussed, that consider cloud system dynamics, interoperability and heterogeneity issues. The intention of the authors is to elicit the characteristics of a given HPC code and to produce from that information a model that reflects the resource requirements of the HPC job in so-called cooperative e-science infrastructures. In [9], it is reported that schedulers in Hypervisors, such as in XEN, cannot handle adequately heterogeneous workloads from high and low performance compute jobs at the same time. The reasons for that is that inter-VM communication inside of the same HPC job is degraded by the fact that a VM can be descheduled in the very moment of their communication with another VM that is not descheduled. As a result, both cannot proceed. The authors suggest to schedule HPC jobs not on a cloud-wide basis, but only inside of isolated subsets of cloud resources to limit inter-job interferences. The authors are using a predictor model and a software implementation of it for a prognosis, which VM will communicate with which other, in order to avoid descheduling at the wrong point in time. Furthermore, they migrate a communication-intensive group of VMs to another resource subset, in order to make IO-dominant HPC more effective. This is accomplished by a scheduler that is aware of the IO activities of VMs, i.e. of their ongoing communication relationships.

**VM Placement:** In article [4], the idea of a better placement of VMs by Nova is deepened. The authors have modified Nova, in order to make it aware of the underlying cloud hardware, of the topology of the interconnect network, of the arrangement of resources and of interferences between jobs because of noisy neighbours.

**Throughput vs User Satisfaction:** In [10], it is reported about a distributed job management system that is supposed to support millions of small HPC jobs. This system aims to the big commercial CSPs such as Amazon and Google. The focus lies on high throughput and good utilization, which is exactly what CSPs need, but not in minimizing the elapsed time for individual HPC jobs. In [11], the term HPC-as-a-Service is introduced as a new offer from CSPs.

The project tries to bridge the gap between what a CSP can offer as compute resources to his clients at a specific moment in time and for a specific price, and what clients want to have at that moment. It is explained that both sides (clients and CSPs) exhibit big variances and heterogeneities. It is furthermore pinpointed that a multi-criteria optimization of cloud resources is needed because of that. The authors used for that purpose mixed integer linear programming and a stochastic optimization model for efficient HPC resource sharing for service provisioning. Their focus lies on the cost-benefit of cloud resources.

**Trustfulness of Results:** The papers we have reviewed have contributed to HPC-efficient clouds. However, some papers were using not a real cloud, but some simulator, or they have not used a real HPC code, but synthetic load generators. From our point of view, it was not always clear how realistic the achieved results are. Therefore, we followed the path of real hardware executing a widely-used HPC package, and to make with this package real measurements on a standard cloud in the hope to achieve more realistic results.

### 3. PROJECT DESCRIPTION

For our project, we built an own cloud, installed OpenStack and used OpenFOAM as HPC benchmark, which is based on the MPI parallelization standard [13]. Furthermore, shell scripts were written to automate the OpenFOAM benchmarks by running them with various parameters and set-ups.

#### 3.1 Easier Possibilities

Before we started to establish an own cloud, we have investigated the subsequent easier possibilities:

- 1) Installing a cloud on a set of VMs (nested virtualization)
- 2) Using a cloud simulator
- 3) Doing all measurements in a commercial cloud or on a University cloud in a computing centre.

All three options were evaluated, and it is explained in the following why we dismissed them all together.

**Nested Virtualization:** We found out that already a single virtualization that is not nested, decreases HPC speedup and efficiency, unless the latest available hardware accelerators for virtualized computation and communication are engaged, which are described in [14], for example. To install OpenStack on a set of VMs in order to create VMs insides of VMs was therefore not an option. **Cloud Simulators:** From the set of easily available open-source cloud-simulators, we started with a closer look to CloudSim [15]. The alternatives we have considered before and dismissed were GreenCloud [16], iCanCloud [17] and an improved version to the MaGateSim simulator, which is described in [18]. GreenCloud and MaGateSim are for energy-saving cloud-computing, which is not in our focus. From our point-of-view, they were too limited for the required performance analyses. Furthermore, iCanCloud helps to predict the trade-off between cost and performance, which is also not relevant for us. A closer look into CloudSim revealed that it contains a very limited model for communication, which cannot reflect sufficiently the MPI-based communication of OpenFOAM, and also not the complex virtualized

communication of OpenStack. For example, it does not model separately inter-core, inter-processor and inter-server communication. CloudSim alone is too restricted for what we need. Additionally, we found the paper [19], which stated that the results from CloudSim are not realistic. Because of that, the authors of [19] created a substantial update called NetworkCloudSim, which supports a more advanced bandwidth/latency model. We concluded that from all network simulators only NetworkCloudSim has relevance for us. It is in principle possible to profit from NetworkCloudSim by a simulative exploration of models for cloud applications. These models are typically defined in terms of estimated job duration and communication times between the parallel tasks of a job. However, the claim of the authors that NetworkCloudSim allows for precise evaluation of scheduling algorithms in scientific, MPI-based applications, including the modelling of a data centre's interconnection network could not be verified by us. The problem with this claim was that the authors did not provide any figures or examples from real measurements to gauge the many NetworkCloudSim parameters. Furthermore, hardware accelerators, such as Single Root IO-Virtualization (SR-IOV) [14], [20], for example, which are nowadays indispensable for an efficient virtualized computation and communication, are not contained in NetworkCloudSim. They must be implemented by the user with high effort. Furthermore, no predefined model exists for inter-VM communication via the KVM Hypervisor or for the OpenVSwitch [21] of OpenStack. Such a model would have been very hard to realize by the user, because the virtual networks in OpenStack follow the principle of Software-Defined Networking (SDN), with the result that they are dynamically variable over time, which is a challenge for every model. Furthermore, NetworkCloudSim does not provide a ready option to model the influence of tunneling protocols, such as GRE [22], and of VLAN or VXLAN [23]. We cannot ignore them, since they are frequently used in clouds. Finally, due to our literature search, nobody else has modelled so far in NetworkCloudSim the distinct configuration parameters of a Hypervisor, of a hardware accelerator, of VLANs or of tunneling protocols. This meant for us that it was not possible to obtain realistic results without tremendous own efforts for software development and parameter gauging: NetworkCloudSim does not provide ready options to model the communication structure and setup of an HPC application reliably enough. This made a real cloud indispensable for us. The only question was, whether an own cloud or an alien cloud would be the easier solution. Existing Commercial or University Clouds: We learned quickly that it is not possible to change on-the-fly the interconnect structures in a commercial or University cloud, or to add contemporary hardware accelerators, because this disturbs productive operation. Furthermore, system administrator rights would have been needed, which cannot be obtained for an alien infrastructure. Further problems have been that no CSP known to us allows for specific placements of user VMs in his computing centre in order to influence deliberately the other loads, which co-exist at the same time. Because of that, it was not possible for us to exclude measurement errors caused by the Noisy Neighbour problem. Neither could we ensure this way the reproducibility of the measurement results. For that reasons, we decided that all three options discussed above are not viable, and we decided to establish an own cloud.

### 3.2 Our Project Cloud

Our cloud consists of 17 used servers from Dell and Sun, which were at the time of the measurements older than 4 years. We had a total number of 76 Cores, 292 GB RAM and 19 TB as Disk Storage. The servers were coupled by 17 Infiniband network interface cards of 40 Gbit/s each and a 40 Gbit/s Infiniband switch. The interfaces are of type Mellanox MHQH19B-XTR and are using QSFP copper cables, as well as the switch itself, which is of type Mellanox

Infiniscale IS5023. The switch has 18 ports and a low port-to-port latency of 100 ns only. In parallel to that high-speed network, a standard communication system was installed, that comprises 17 Ethernet cards of 1 Gbits/s each and a Ethernet switch with 24 port of 1 Gbits/s, respectively, to allow for performance comparisons between the two couplings. The host OS for the cloud was Ubuntu 14.04.01 with the OpenStack IceHouse release installed, while Ubuntu 12.04.05 was used throughout as Guest OS.

### 3.3 Integration of Infiniband in Our Cloud

For the integration of Infiniband in our cloud, the virtual Ethernet-network interface-cards (NICs), which are the standard API of KVM for the user, were realized by us by means of the TAP device driver [24]. TAP simulates a NIC by software, and users communicate via TAP by read/write file operations. These operations are translated by TAP into payloads for virtual Ethernet frames, which are subsequently forwarded by OpenVSwitch by means of L2 switching. OpenVSwitch is as KVM an important component of OpenStack. Both are initialized and configured by OpenStack. After that step, OpenVSwitch provides for every TAP a virtual switch port, at which TAP can feed-in its virtual Ethernet frames. Additionally, OpenStack and KVM provide for the VMs of every customer an own VLAN, which is isolated from the VLANs of other customers, in order to provide for IT security. OpenVSwitch processes each virtual Ethernet frame such that the frame is either delivered by means of the user VLAN to a VM in the same server, or alternatively, such that a switch output-port forwards the frame via a GRE tunneling protocol to another Host OS. Since the transportation of an IP packet in an Infiniband Frame is not possible, the Infiniband-over-IP (IPoIB) protocol [25] was added as carrier. We have found no other possibility to integrate Infiniband in OpenStack without SR-IOV.

### 3.4 Our HPC Application

As an example for a HPC application, the widely used OpenFOAM was selected, which is based on Open MPI [26]. It is a parallel HPC code for the numeric solution of Laplace and Navier-Stokes equations, i.e. for the calculation of laminar and turbulent flows of compressible and incompressible fluids, which are gases or liquids. OpenFOAM has additional solvers for general particle flows, for combustion, molecular dynamics, heat transfer, electromagnetic problems, solid elastic bodies, and other purposes, which were not used by us. The reason for the latter was: before we started with OpenFOAM, we performed a questionnaire by asking users of OpenFOAM what they are exactly doing, and what their expectations are when executing the code on a cloud. From that questionnaire, we understood that OpenFOAM is mainly used to solve the Navier-Stokes equations, and we learned also that users considered OpenFOAM and OpenStack as unfavourable combination, unless the latest server generation is used in the cloud.

## 4. PERFORMANCE TESTS

Initially, we configured OpenFOAM to execute the Dam Brake example that comes with the 2.2.1 distribution, because it is well documented. In this example, there are 7700 grid points for geometric objects in two dimensions. One second in reality is simulated by 1000 time steps. After a first simulation run, we modified the initial configuration to make more advanced tests. All measurements were repeated 50 times, and the first run in each measurement cycle was deleted in order to exclude transient effects.

#### 4.1. Measurement Results

The execution-time results of all test set-ups are shown in table 1. The results were post-processed by calculating the speed-up and the efficiency of the cloud. These two metrics are defined by:

Definition 1. Speed-up  $S$  is the ratio of the execution times of a sequential code before and after virtualization and parallelization.

Definition 2. Efficiency  $E$  is the utilization of  $n$  server cores or OpenStack vCPUs and defined as  $E = S/n$ .

Table 1. Measurement results for the set-ups 1-7.

Set-Up	Wall-Time [s]	Speed Up	Efficiency [%]
1a: 1 core, bare metal	144	1	100
1b: 4 cores, 2 CPUs, 1 server, bare metal	46	3.1	78
2a: 1 core, 1 KVM	180	0.8	80
2b: 4 cores, 2 CPUs, 1 server, 1 KVM	62	2.3	58
3: nested virtualization, 1 core, 2 KVMs	-	-	-
4a: 1 core, 1 KVM, OpenStack	154	0.94	94
4b: 4 cores, 2 CPUs, 1 server, 4 KVMs, OpenStack	60	2.4	60
5: 4 cores, 4 CPUs, 4 servers, 4 Ethernets, 4 KVMs, OpenStack	320	0.45	11
6: 16 cores, 4 CPUs, 4 servers, 4 Ethernets, 16 KVMs, OpenStack	670	0.21	5
7a: 4 cores, 4 CPUs, 4 SUN servers, 4 Infinibands, 4 KVMs, OpenStack	237	0.61	15
7b: 16 cores, 4 CPUs, 4 SUN servers, 4 Infinibands, 16 KVMs, OpenStack	998	0.14	4

#### 4.2 Evaluation of Measurement Results

According to set-up 1a, the reference execution time for all subsequent measurements was determined as 144 s. This value indicates that the problem size compared to usual HPC execution times is too small, although it is the standard example of OpenFOAM. From set-up 1b, it can be seen that parallelization is beneficial in our cloud if the code execution takes place in the same server. However, efficiency already drops by 22 percentage points to 78% on 4 cores. On a

supercomputer or a parallel computer, OpenFOAM should scale well until about 1000 cores, according to its manual. A drop of 22 points already at 4 cores is a sign that the communication time cannot be neglected compared to the computation time. It confirms that the used problem size is too small. From set-up 2a, it can be seen that virtualization causes the efficiency to drop by 20 percentage points. This can be explained by the fact that only the relatively old AMD V [14] accelerator in the CPUs was used, but not newer methods, which could reduce better the efficiency losses that are caused by virtualization. Set-up 2b shows that the simultaneous usage of virtualization and parallelization reduces efficiency by 42 percentage points, which could be expected already by adding the figures 1b and 2a. Set-up 3 was not possible to conduct, because the VM that was created by KVM inside of another VM -according to nested virtualization was not able to run its guest OS. The reason for this is unknown. Set-up 4a shows that the cloud OS incurs an overhead, such that the efficiency drops by 6 percentage points which is low. However, Set-up 4b shows an efficiency drop of 40 percentage points for the parallel code. In Set-up 5, a drastic drop down to 11% can be observed in case of code distribution over 4 vCPUs, which are residing on 4 different servers, which is not tolerable for HPC. In set-up 6, the situation escalates to 5% efficiency, when the code is executed in parallel on 16 vCPUs from 4 servers. Finally, the biggest surprise to us was the measurement in setup 7a, because efficiency increased only to 15%, although Infiniband with the 40-fold data rate was used instead of 1 Gbits/s Ethernet. In setup 7b, Infiniband is even worse than Ethernet in setup 6, which is remarkable. In the following, it is explained how this has happened.

### 4.3 Communication Overheads

We explain the surprising behaviour of Infiniband by the following facts: 1.) the payload of the Infiniband network in setup 7b is shorter than in setup 7a, because the same problem size is divided by 16 vCPUs instead of 4. Shorter payloads, however, increase the Host OS overhead and thus decrease efficiency, because the Infiniband header remains the same. 2.) the minimum transport unit Infiniband can carry is 256 Bytes, while Ethernet needs only 64 Bytes. However, as soon as the problem size gets too small, not enough intermediate computational results can be exchanged between grid borders. As a consequence, more padding bytes are needed in case of Infiniband than for Ethernet. 3.) Infiniband was integrated by us by means of several additional device drivers and protocols, because without SR-IOV there was no other possibility. As a consequence, two VMs, which are located on different servers, communicate with each other by means of payloads in virtual Ethernet frames, which are transmitted via Berkeley Sockets and TCP/IP in the Guest OS. In the Host OS, we used TAP [24], OpenVSwitch, GRE, IP, IPoIB, TUN, and OFED verbs [27]. This creates significant overhead.

## 5. RESULTS AND CONCLUSIONS

The measurements in our cloud have shown that under the given hardware and software configuration the best speed-up and efficiency could be achieved, if the code was executed by the cores of one CPU or by the CPUs of one server. This is explained by the multiple overheads for virtualized communication that are involved otherwise. Furthermore, the measurements have shown that it is not sufficient to replace the Ethernet network in a standard cloud by Infiniband. Other improvements must be added as well, otherwise a 40 Gbits/s Infiniband can be even slower than a 1 Gbits/s Ethernet. Furthermore, in case of HPC code distribution to different cloud servers it was not possible for the servers to compensate for the resulting communication overhead,

because modern hardware accelerator for virtualized communication, such as SR-IOV, were not used by us. Because of that, vCPU data multiplexing and VM data switching was accomplished by OpenVSwitch. The communication overhead made it also impossible to use the remote DMA feature of Infiniband, because multiple extra protocols, device drivers and interfaces had to be added. Our first conclusion is that it is not possible to use clouds with old servers or with standard servers that do not include the latest hardware accelerators for both, virtualized computation and communication. Otherwise OpenStack +KVM +OpenVSwitch is no efficient combination. According to literature and our own findings, there are some cloud-intrinsic problems that make HPC potentially HPC-inefficient. These problems are: 1.) existing cloud schedulers ignore the needs of HPC tasks. One example for that is that heterogeneous cloud hardware and software with different performance capabilities can easily be coupled in a cloud, but with the consequence that the scheduler allocates parallel subtasks of the same HPC job to hardware of different performance, although the subtasks have the same computational intensity. Another example is that communicating subtasks can be scheduled at different points in time, thereby disabling efficient rendezvous-based communication. A third example is that we could not find-out, whether KVM memory protection provides for fast inter-vCPU and for inter-VM communication. However such a communication is indispensable for Open MPI data exchange via shared cache or share memory. 2.) Multi-tenancy and its consequence, the noisy-neighbour problem, effects that the vCPUs of the same VM and that the VMs of the same parallel job are competing with each other for cloud resources. Or second conclusion is therefore that the improvements, which are listed below, should be added to OpenStack to achieve HPC efficiency. Or third conclusion is that it is also not sufficient to install OpenStack on an existing parallel computer, in order to obtain the benefits clouds have. This will result in faster job execution, but efficiency problems were still not solved. A way-out is what most HPC Cloud providers, such as Nimbix, Sabalcore and Nephoscale are doing: they run HPC load on bare metal and with Infiniband. In addition, companies like the UberCloud are using Docker containers [28] and enhancements of it, which are as flexible as VMs, but very lightweight, in order to reduce overhead. Furthermore, customers can use bigger servers in the cloud with more cores and more main memory as they have at home, and thus run their code faster on the one CPU or on one server as at home.

## 6. OUTLOOK

It is our working hypothesis that it is possible to turn every standard cloud into an HPC-capable tool, provided that several or all of the subsequent improvements are made. Of course, not all of them are new, but nevertheless important, which is why we have listed them below.

1. Replacement of Standard Ethernet for inter-server communication by fast Ethernet or Infiniband of at least 10 Gbit/s (better 40 Gbit/s) that is driven by 8-lane PCIe interfaces as minimum.
2. Engagement of the latest hardware accelerators for virtualized communication and computation not only in the CPU, but also on the server motherboard and in the PCIe peripherals. The accelerator that is mandatory as minimum equipment is either SR-IOV or VT-d [29]. However, both do not come for free and have disadvantages as well with respect to system administration and IT security: they do not allow splitting the real interconnection network of the cloud into separate tenant VLANs or VXLANs. A workaround is to use Infiniband partitions. Unfortunately, partitions cannot be configured by OpenStack Neutron.

3. Proper configuration and activation of BIOS-, Hypervisor and Cloud OS options and flags to fully exploit the aforementioned accelerators. In practice, this can be quite difficult.
4. Avoidance of memory and core over committing by too much virtualization. The amount of VM launches and virtual main memory should be limited and carefully monitored. This prevents from excessive paging in guest and host OSes.
5. Avoidance of nested virtualization, unless proper accelerators are available.
6. Employment of Open MPI because of its efficiency, its automatic selection of the fastest communication paths between processing elements and its reluctance in using TCP/IP.
7. Replacement of TCP/IP in the Guest OSes and of IP in the Host OS by stubs such as the Mellanox Messaging Accelerator MXM [14], [30] or by Myrinet Open-MX [31]. This is possible because OpenStack replaces L3 routing by L2 switching, as long as the cloud is in the same hall or rack.
8. Replacement of the existing Guest OS und Host OS Schedulers by an approach that allows for priorities and that is aware of the cloud hardware, its network topology and the problem noisy neighbours, and that differentiates jobs into classes ranging from low performance to high performance computing.
9. Adding to the cloud OS scheduler a gang scheduling in space that allocates communicating vCPUs and VMs to cores in the same CPU chip or to cores in the same server, in order to improve inter-process communication.
10. Adding to the cloud OS scheduler a gang scheduling in time that that schedules communicating vCPUs simultaneously, in order to allow for rendezvous Host OS via blocking Send/Receive.
11. Adding to the cloud OS scheduler a gang scheduling in capability: IO-intensive HPC jobs should be scheduled to servers with fast peripherals. To accomplish that, a job control language (JCL) is needed for clouds, which informs the scheduler about IO-intensiveness of the jobs.
12. Avoidance of waiting for IO resources by an advanced reservation of peripherals, that should be implemented already in the jobs JCL suggested previously.
13. Enhancement of the cloud OS scheduler by a performance delivery model of the cloud and by a resource consumption model of its clients for predictive resource planning. These demand/offer models should allow for interconnect topology-awareness, server hardware-awareness and application awareness and contribute to more profound scheduling decisions.
14. Avoidance of the noisy-neighbour problem by partial batch processing in the cloud. The cloud OS scheduler should make an exclusive allocation of vCPUs to jobs, without any time sharing of cores between jobs and users. The time sharing in the cloud should be restricted to another subset of cores, because it disturbs inter-VM and inter-vCPU communication.
15. Incrementing of the problem sizes to at least 100 000 grid points to make the ratio between computation and communication better.
16. Integration of Infiniband management into OpenStack Neutron to avoid alien tools.
17. Provisioning of two different Infiniband device driver, one that is optimized for short L2 frames, the other that is optimized for long L3 IP V6-Jumbo packets: this allows to adapt to the communication requirements of the various VMs.

## ACKNOWLEDGEMENT

The project was funded by the Scientific Simulation centre Clausthal-Goettingen (SWZ) under contract #11.4.1.

## REFERENCES

- [1] U.S. Department of Energy, Office of Advanced Scientific Computing Research (ASCR), The Magellan Report on Cloud Computing for Science, December, 2011.
- [2] A. Gupta, Techniques For Efficient High Performance Computing In The Cloud, Dissertation, University of Illinois at Urbana-Champaign, 2014.
- [3] A. Gupta, L. V. Kale, Towards Efficient Mapping, Scheduling, and Execution of HPC applications, 27th IEEE International Symposium on Parallel & Distributed Processing Workshop, May 20-24, Boston, USA, 2013.
- [4] A. Gupta, L. V. Kale, D. Milojicic, P. Faraboschi, S. M. Balle, HPC-Aware VM Placement in Infrastructure Clouds, IEEE International Conference on Cloud Engineering, March 25-28, San Francisco, USA, 2013.
- [5] OpenStack, <http://www.openstack.org/>, retrieved at 19.06.2015.
- [6] OpenFoam, <http://www.openfoam.com/>, retrieved at 19.06.2015.
- [7] Stelios Sotiriadis, Nik Bessis, Fatos Xhafa, Nick Antonopoulos, From metacomputing to interoperable infrastructures: A review of metaschedulers for HPC, grid and cloud, 26th IEEE International Conference on Advanced Information Networking and Applications, March 26-29, Fukuoka, Japan, 2012.
- [8] N. Bessis, S. Sotiriadis, V. Cristea, F. Pop, Modelling Requirements for Enabling Meta-Scheduling, Third International Conference on Intelligent Networking and Collaborative Systems, 30 Nov 02 Dec., Fukuoka Institute of Technology, Fukuoka, Japan, 2011.
- [9] Yanyan Hu, Xiang Long, Jiong Zhang, Enhance Virtualized HPC System Based on I/O Behavior Perception and Asymmetric Scheduling, 14th IEEE International Conference on High Performance Computing and Communications, 25-27 Jun, Liverpool, UK, 2012.
- [10] Iman Sadooghi, Sandeep Palur, Ajay Anthony, Isha Kapur, Karthik Belagodu, Pankaj Purandare, Kiran, Ramamurty, Ke Wang, Ioan Raicu, Achieving Efficient Distributed Scheduling with Message Queues in the Cloud for Many-Task Computing and High-Performance Computing, 14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, May 26-29, Chicago, USA, 2014.
- [11] Han Zhao, Xiaolin Li, Designing Flexible Resource Rental Models for Implementing, HPC-as-a-Service in Cloud, 26th IEEE International Parallel and Distributed Processing Symposium Workshops & PhD Forum, DOI 10.1109/IPDPSW.2012.324, 21-25 May, Shanghai, China, 2012.
- [12] R. Ledyayev, H. Richter, High Performance Computing in a Cloud Using OpenStack, The Fifth International Conference on Cloud Computing, GRIDs, and Virtualization, CLOUD COMPUTING 2014, <http://www.iaria.org/conferences2014/CLOUDCOMPUTING14.html>, Venice, Italy, 6 pages, May 25 29, 2014.
- [13] MPI, <http://www.mcs.anl.gov/research/projects/mpi/>, retrieved at 19.06.2015
- [14] H. Richter, A. Keidel, Hochleistungsrechnen und Echtzeit in virtualisierten Maschinen und Clouds Die Intel Virtualisierungshilfen, in IfI Technical Report Series ISSN 1860-8477, IfI-14-03, <http://www.in.tuclausthal.de/forschung/technical-reports/>, editor: Department of Computer Science, Clausthal University of Technology, Germany, 44 pages, 2014.
- [15] CloudSim, <http://www.cloudbus.org/cloudsim/>, retrieved at 19.06.2015.
- [16] GreenCloud, <http://www.opensourceforu.com/2015/01/getting-startedgreencloudsimulator/>, retrieved at 19.06.2015.
- [17] iCanCloud, <http://www.arcos.inf.uc3m.es/icancloud/Home.html>, retrieved at 19.06.2015.
- [18] C. Thiam, G. Da Costa, J.-M. Pierson, Cooperative Scheduling Anti-load balancing Algorithm for Cloud : CSAAC, IEEE International Conference on Cloud Computing Technology and Science, 2-5 Dec., Bristol, UK, 2013.

- [19] S. K. Garg, R. Buyya, NetworkCloudSim: Modelling Parallel Applications in Cloud Simulations, Proc. Fourth IEEE International Conference on Utility and Cloud Computing, 5-8 Dec., Melbourne, Australia, 2011.
- [20] SR-IOV, <https://msdn.microsoft.com/en-us/library/windows/hardware/hh440148%28v=vs.85%29.aspx>, retrieved at 19.06.2015.
- [21] openvswitch, <http://openvswitch.org/>, retrieved at 19.06.2015.
- [22] GRE, <https://tools.ietf.org/html/rfc2784>, retrieved at 19.06.2015.
- [23] VXLAN, <https://tools.ietf.org/html/rfc7348>, retrieved at 19.06.2015.
- [24] tuntap device driver, <https://www.kernel.org/doc/Documentation/networking/tuntap.txt>, retrieved at 23.06.2015.
- [25] IPoIB, <https://www.kernel.org/doc/Documentation/infiniband/ipoib.txt>, retrieved at 23.06.2015.
- [26] Open MPI, <http://www.open-mpi.org/>, retrieved at 23.06.2015.
- [27] OFED, [http://www.mellanox.com/related-docs/prod software/Mellanox OFED Linux User Manual v3.1.0.1.pdf](http://www.mellanox.com/related-docs/prod%20software/Mellanox%20OFED%20Linux%20User%20Manual%20v3.1.0.1.pdf), retrieved at 23.06.2015.
- [28] Docker Container, <https://www.docker.com/>, retrieved at 23.06.2015.
- [29] Intel VT-D, <https://software.intel.com/en-us/blogs/2009/06/25/understanding-vt-d-intel-virtualization-technology-for-directed-io>, retrieved at 19.06.2015.
- [30] Mellanox Technologies, Mellanox Messaging Library UserManual, [http://www.mellanox.com/page/products\\_dyn?product\\_family=135&menu\\_section=73](http://www.mellanox.com/page/products_dyn?product_family=135&menu_section=73), retrieved at 24.07.2014.
- [31] Inria Open-MX, Open-MX Myrinet Express over Generic Ethernet Hardware, <http://openmx.gforge.inria.fr>, retrieved at 19.06.2015.

## AUTHORS

**Harald Richter** was born in Stuttgart/Germany in 1958. In 1983, he got a ‘Dipl.-Ing.’ diploma degree in Electrical Engineering with specialisation in Computer Engineering from the University of Stuttgart. In 1988, he received a ‘Dr.-Ing.’ degree in Electrical Engineering from Munich University of Technology, and in 1998, he acquired a ‘Dr. rer.nat.habil.’ degree in Computer Engineering from the same University. Since 2000, he has the chair of Technical Informatics and Computer Systems at Clausthal University of Technology, where he works until today. He teaches computer organization and computer networks. His research interests are Real-Time Communication in Computer Networks, Real-Time Data Acquisition, High-Performance Computing and Simulation and Renewable Energies .



*INTENTIONAL BLANK*

# REFERENCE ARCHITECTURE FOR SMAC SOLUTIONS

Shankar Kambhampaty<sup>1</sup> and Sasirekha Kambhampaty<sup>2</sup>

<sup>1</sup>Computer Science Corporation (CSC), India  
skambhampaty@gmail.com

<sup>2</sup>Student, Department of Computer Science, GRIET, Hyderabad, India  
ksasirekha@gmail.com

## **ABSTRACT**

*Web and internet computing is evolving into a combination of social media, mobile, analytics and cloud (SMAC) solutions. There is a need for an integrated approach when developing solutions that address web scale requirements with technologies that enable SMAC solutions. This paper presents an architecture model for the integrated approach that can form the basis for solutions and result in reuse, integration and agility for the business and IT in an enterprise.*

## **KEYWORDS**

*Architecture, Model, Web-scale, Social Media, Mobile, Analytics & Cloud Computing*

## **1. INTRODUCTION**

The rise of mobile apps has been causing increasing demand for accessing functionality from outside the infrastructure of enterprises. Consumers are demanding information relevant to their context anywhere, anytime, on any device. At the same time, the rise of “Internet of Things” also requires a standard interface for devices to communicate the data captured by different types of devices (such as refrigerators & dish washers) through published interfaces. Additionally, social media applications (Facebook, Twitter etc.) are fast becoming access channels for consumers to interact with products (e.g. Salesforce CRM) and services (such as internet banking) of enterprises. All these taken together constitute web-scale demand that needs to be addressed by solutions of enterprise.

Web and internet computing technologies are making significant advances with evolving technologies that may be grouped under four categories – social media, mobile, analytics and cloud computing [1]. Enterprises employ these technologies to deal with the web-scale demand when offering products and services to their customers.

The problem, however, is that point solutions are being provided in several enterprises without taking a holistic view of the current and future needs of the enterprise leading to “patch work” of the IT in the enterprise thereby increasing the cost in the long run.

The purpose of this paper is to provide reference architecture to address the above problem and enable in reuse, integration and agility thereby reducing the total cost of ownership for the enterprise.

## 2. WEB-SCALE DEMAND

“Web-scale describes the tendency of modern architectures to grow at (far-) greater-than-linear rates. Systems that claim to be Web-scale are able to handle rapid growth efficiently and not have bottlenecks that require re-architecting at critical moments” [2].

Some of the sources of web-scale demand are as follows:

- 1) Data from sensors (such as electricity meters) to applications that manage business processes (e.g. power load management).
- 2) Consumer devices (such as wearables) that provide data on continuous basis to remote healthcare monitoring systems.
- 3) Data feeds from social media sites that provide insights on consumer buying behaviour.
- 4) Mobile devices accessing news sites when sensational events happen round the world.

Engineering firms that deal with large number of sensor data and enterprises that provide products and services to a substantial volume of consumers are likely to experience the web-scale demand. Gartner predicts that “By 2017 Web-Scale IT Will Be an Architectural Approach Found Operating in 50 Percent of Global Enterprises” [3]. This is because of the disruption in how, traditionally, business has been conducted and the trend towards digital business models.

## 3. SMAC SOLUTIONS

Several key technology advancements have been taking place in social media, mobile, analytics and cloud computing that are collectively referred to by the term SMAC.

Social media platforms are being leveraged by enterprises as one of the access channels for customers. For instance retail banks (such as ICICIBank) provide apps on social media platforms (e.g. Facebook) to allow customers to perform the day-to-day banking transactions. This requires business solutions of the enterprise (e.g. core banking solution) to be integrated with the social media platform to enable large number of transactions to flow through the social media channel [4]. Social media platforms being open to all, can potentially result in millions of transactions that have to be handled in a secure manner over relatively short periods of time.

Mobile platforms are presenting both a huge opportunity and significant challenge to large enterprises. It is necessary to provide mobile apps to enterprise users and customers to interact with the enterprise for products and services while supporting a wide variety of mobile operating systems (iOS, Android, Windows etc.) that run on a multitude of devices with varying form factors. The mobile apps have to be continuously kept up-to-date and have to interact with the core business solutions of the enterprise in a secure and a user-friendly manner. The back-end solutions for the mobile enables need to be highly scalable, insulated against the regularly

changing front-end mobile apps. An MBaaS platform, mobile backend as a service, may also be part of solution mix to support integration and address the mobile notification requirements of the enterprise [5].

Analytics solutions typically extract intelligence from structured and unstructured data to help enterprises make business decisions. The business enterprises require real-time decisions to be made based on the large amount of business and transaction data that flows from the social media and mobile applications. Big data solutions based on technologies such as Storm and Hadoop are central to organization IT landscape for this purpose [6].

Cloud platforms and related solutions provide the back-end infrastructure to host and deploy applications that support business in addition to traditional IT within enterprises. The cloud platforms are enabling scalability and reducing total cost of ownership (TCO) through pay-per-use cost model while providing one or more of the following “as a service” offerings [7]:

- Infrastructure as a Service (IaaS)
- Platform as Service (PaaS)
- Software as a Service (SaaS)

#### **4. NEED FOR INTEGRATED APPROACH**

In most enterprises, the solutions for each of the SMAC technologies are being led and managed by different teams from various units without taking into account an enterprise-level view. Often, this is because the driving factors for the solutions for each of the SMAC technologies are different and justification for business investments and return on investment (ROI) are measured differently. While an enterprise architecture view may exist in many enterprises and SMAC technologies may be formally be identified as strategic in their roadmaps, the fact that they are managed by different groups in the enterprise with goals that are not necessarily aligned does not foster an integrated model for the SMAC solutions. Consequently, SMAC solutions in enterprises have evolved to be disparate systems with little or no integration resulting in efficiencies, increased total cost of ownership (TCO) and lesser time to market [8].

An integrated approach with a common layer acting as “broker” for social media, mobile, analytics and cloud platforms can enable scalable (due to cloud support), customer facing (due to social media and mobile touch points) and intelligent (drawing from analytics system interfaces) solutions. An example of such a solution is a travel solution that gathers intelligence related to customer buying behaviour from social media and airline web sites and pushes alternate or related product purchasing options to clients through mobile apps. Another example is a power company that switches users to different power providers based on load, time of the day and pricing using data from sensors in the power meters and informing users through mobile text messages.

#### **5. SERVICES MODEL – THE FOUNDATION**

The services model that provides loose coupling of service consumers with service providers and integrated with a “broker” pattern, provides a good foundation for an integrated architecture model for SMAC solutions [9]. Equally important is the fact that most enterprises with mature IT

practices have made substantial investments in implementations of Service-Oriented Architecture (SOA).

A generic services model for an enterprise may be defined based on four types of services, namely, *activity services (A)*, *business process services (B)*, *client services (C)* and *data services (D)* [10]. Figure 1 depicts such an enterprise-level generic architecture based on the four types of services.

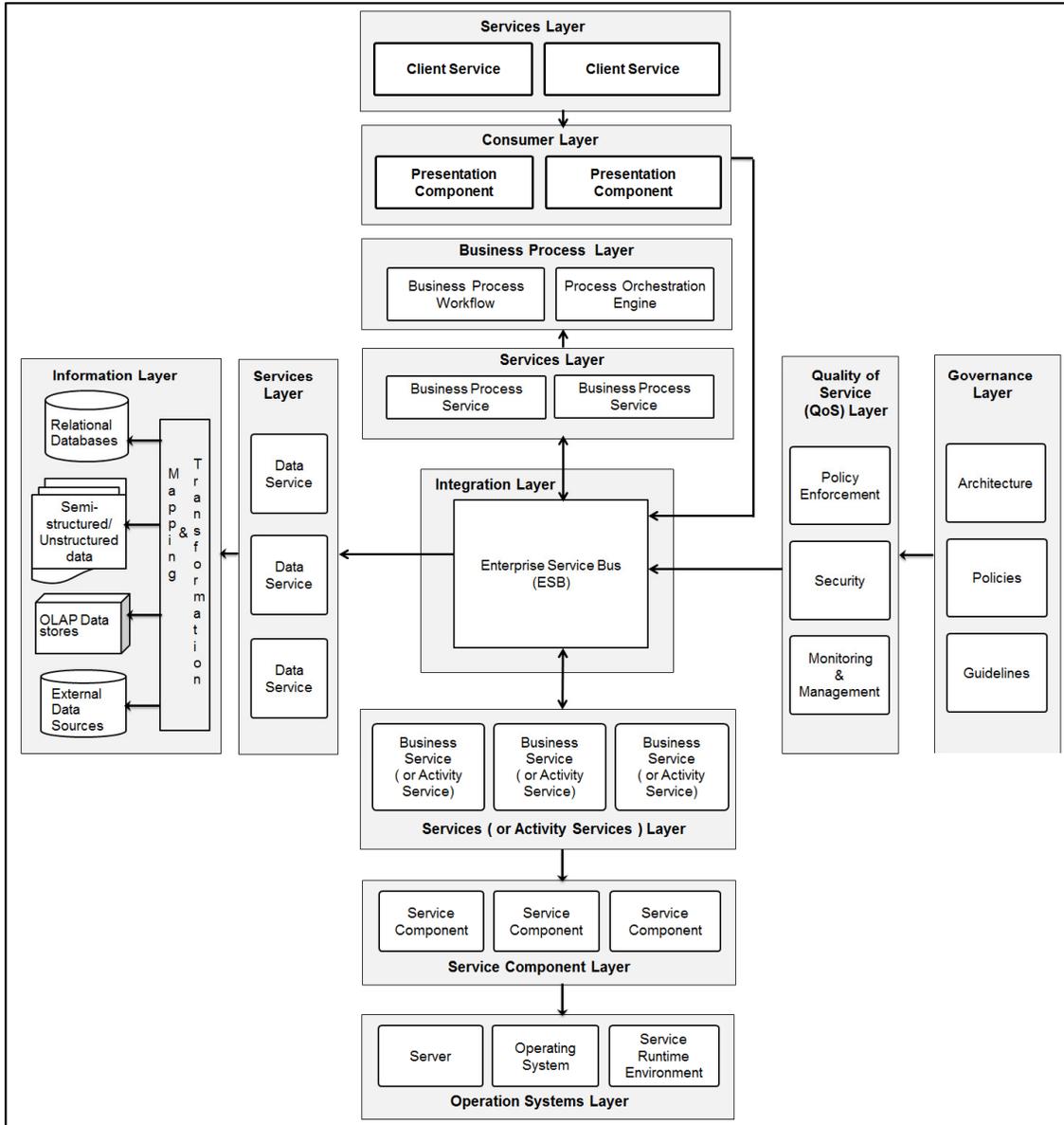


Figure 1: Enterprise-level generic architecture based on services model

The generic architecture shown in Figure 1 represents a logical perspective of the four types of services and their relationships.

A description of the services is as follows:

1. **Activity services (or Business services):** The activity services encapsulate functionality exposed as services in various business applications. These services are reusable business level services that can be orchestrated as part of a configured business process. These services will be referred to as type “A” services.

2. **Business process services:** Business process services handle workflow requirements (e.g. approval processes) through orchestration of business processes, each step of which is implemented as a service. These services enable externalization of business processes through their specification as workflows and orchestration by process orchestration engine resulting in agility for the enterprise. They will be referred to as type “B” services.

3. **Client services:** Client services provide content meant for “front-end” applications of the enterprise to enable clients/partners to access their business data (e.g. order information) and enterprise users to get an aggregated view (e.g. enterprise dashboard). APIs defined by an enterprise for consumption by clients/partners may be implemented with this category of services. These services will be referred to as type “C” services.

4. **Data services:** Data services provide access to data in various sources. There are typically two types of data services that have to be provided in an enterprise:

- Structured data stores (e.g. relational databases)
- Unstructured data stores (e.g. Big data data stores that run analytics)

These services will be referred to as type “D” services.

The Enterprise Service Bus (ESB) in the integration layer enables a smooth communication between the service providers and service consumers. An Enterprise Service Bus (ESB) pattern abstracts the mediation and interaction elements needed for communication on a bus that is used for integration of services.

The generic architecture in Figure 1 also addresses the need in enterprises to define policies for key non-functional requirements and their enforcement through elements in governance and quality of service (QoS) layers:

1. Governance – Service interaction (service calls from service consumers to service providers) has to be governed. To make this happen, policies are defined in governance layer for non-functional requirements such as security taking into account key performance indicators for business and SLAs for IT.
2. Quality of Service (QoS) – Service interaction has to be monitored and secure and in accordance with the policies defined. This is ensured through monitoring of service interactions (application monitoring, business activity monitoring and IT systems monitoring) and enforcing of policies at runtime.

## 6. REFERENCE ARCHITECTURE FOR SMAC SOLUTIONS

The emergence of APIs as human readable, externally facing, light-weight services has made APIs central to solutions based on SMAC technologies. In fact, the one aspect common to social media, mobile, analytics and cloud solutions is access of functionality through APIs exposed by the respective platforms. Social media sites (such as Twitter) provide APIs for enabling solutions that respond to user posts (or tweets). Mobile apps consume content by making calls to APIs exposed by back-end applications in the enterprise. Big Data solutions are headed towards providing “intelligence as a service” for use by applications across the enterprise. Thus, APIs can serve to be the “glue” for the integrated architecture model.

With the APIs gaining rapid adoption, technologies that support API solutions such as API Gateways have also come into existence that enable integration and play the role of a “broker” for solutions based on SMAC technologies.

Building on the services model based generic architecture, described in the earlier section, and applying the considerations of web scale demand and trends in SMAC technologies, Reference Architecture may be defined for SMAC solutions as shown in Figure 2.

The Reference Architecture for SMAC solutions depicted in Figure 2 brings together the social media, mobile, analytics and cloud solution components through exposing their functionality through services, integrating them through integration layer and orchestrating them through the business process layer.

The consumer layer in Figure 2 shows social media client applications and mobile apps making API calls to the integration layer. *API Gateway* and *Mobile Backend as a Service (MBaaS)* platform are key elements of the integration layer. API Gateway is a “broker” pattern that exposes a standard set of APIs defined to support both intranet applications as well social media and mobile clients. SMAC solutions expose light-weight APIs (which are RESTful web services) instead of SOAP based web services and thus the API Gateway can effectively take the place of an ESB. Enterprises that have significant number of SOAP based web services (or non-customer facing applications) and have made substantial investments in ESB infrastructure may continue to use the ESB for service integration. But given the trend in the industry, the API Gateway is expected to play a key role in integration of SMAC solutions. Another important element in the integration layer is the MBaaS platform that handles the needs of social media and mobile clients, especially by providing push notifications to them. The webscale demand generated by social media and mobile clients are handled by the combination of MBaaS and API Gateway elements in the integration layer.

Analytics applications are part of the information layer and Figure 2 shows data services being exposed to support the requirements of applications in the enterprise. Of particular relevance in this context are the Big data based analytics applications that will see widespread adoption in future. The webscale demand by social media and mobile clients may generate large volume of data relevant to analytics that may be captured by Big data applications as the API invocations are made to the integration layer. The analytics applications enable business intelligence decisions to be made and data services provide the means of querying the analytics applications through service calls in order to make business decisions.

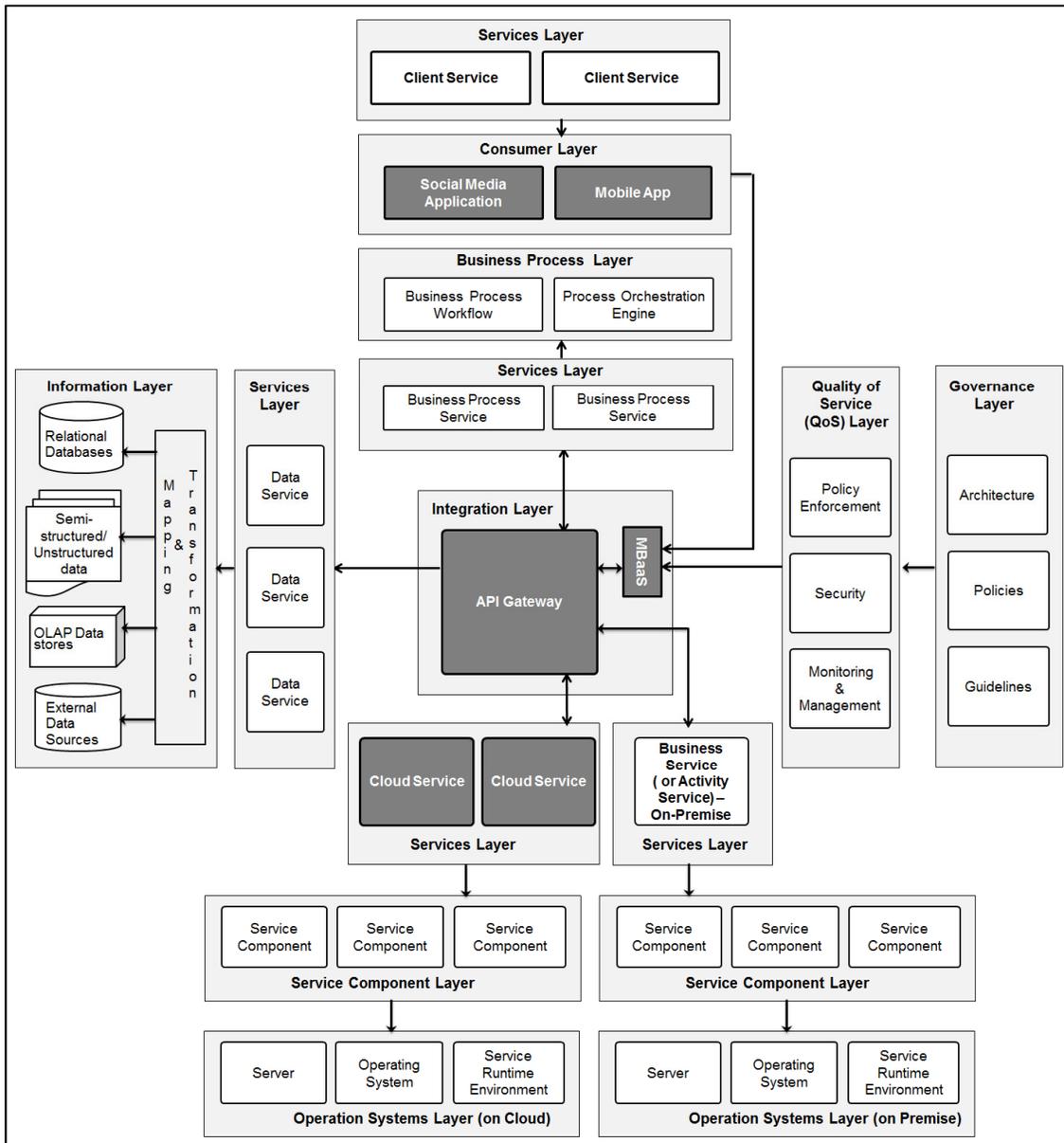


Figure 2: Reference Architecture for SMAC solutions

Enterprises are increasingly expected to move towards hybrid cloud model in future. This would mean a certain amount of business functionality would be implemented in cloud applications and the rest in on-premise applications. Figure 2 shows elements to support both cloud and on-premise applications and their services.

The API Gateway forms the “binding glue” bringing all the services/APIs together to support improved agility and dynamism and reducing the total cost of ownership through reuse and integration [11].

The Reference Architecture so defined provides a consistent and integrated approach to addressing the needs of SMAC solutions.

## **7. IMPLEMENTATION OF REFERENCE ARCHITECTURE**

The Reference Architecture for SMAC solutions is best suited to be implemented at enterprise-level. To that end, the architecture groups within enterprises may consider its adoption as part of their enterprise architecture initiatives.

The following steps are recommended in implementing the Reference Architecture:

- 1) Review of the IT Roadmap of the enterprise.
- 2) Determination of strategic importance of SMAC technologies in meeting the organization goals.
- 3) Assessment of maturity of implementation of services model within the enterprise.
- 4) Identification of products to implement API Gateway and MBaaS platforms.
- 5) Development of business case to justify the investment required for implementation.
- 6) Execution of pilot to provide the technology choices and ROI.

## **8. CONCLUSIONS**

SMAC technologies are seeing increased adoption in web and internet computing. On account of webscale demand and changing needs of organizations, an integrated approach is needed in architecting and implementing SMAC solutions. The services model provides the right foundation for defining a Reference Architecture with an integrated approach. The Reference Architecture proposed in the paper can enable reuse, agility and integration to reduce the total cost of ownership and provide a competitive Enterprise IT architecture to an organization.

## **REFERENCES**

- [1] Evans, N (2013) SMAC and the evolution of IT, Computerworld  
<http://www.computerworld.com/article/2475696/it-transformation/smac-and-the-evolution-of-it.html>
- [2] Leibovici, A (2014) Understanding Web-Scale Properties, Nutanix,  
<http://www.nutanix.com/2014/03/11/understanding-web-scale-properties>
- [3] Gartner (2014) Press Release, Gartner,  
<http://www.gartner.com/newsroom/id/2675916>
- [4] ISACA (2010) Social Media: Business Benefits and Security, Governance and Assurance Perspectives, ISACA,  
<http://www.isaca.org/groups/professional-english/security-trend/groupdocuments/social-media-wh-paper-26-may10-research.pdf>

- [5] Pezzini, M., Guttridge, K., Thoo, E., Thomas, A. (2015) Support Multiple Integration Patterns in Your Mobile App Architecture Strategy, Gartner Document ID: G00270938, [http://www.gartner.com/resources/270900/270938/support\\_multiple\\_integration\\_270938.pdf](http://www.gartner.com/resources/270900/270938/support_multiple_integration_270938.pdf)
- [6] Wähler, K. (2014) Real-Time Stream Processing as Game Changer in a Big Data World with Hadoop and Data Warehouse, InfoQ article, <http://www.infoq.com/articles/stream-processing-hadoop>
- [7] Kambhampaty, S (2010), Service Oriented Architecture for Enterprise and Cloud Applications, 2nd Edition, Wiley India publication, ISBN: 978-81-265-1989-7, <http://www.amazon.in/Service-oriented-Architecture-Enterprise-Cloud-Applications/dp/8126519894>
- [8] Overby, S (2014), How to Tame Social, Mobile, Analytics and Cloud Multisourcing, CIO, <http://www.cio.com/article/2451068/outsourcing/how-to-tame-social-mobile-analytics-and-cloud-multisourcing.html>
- [9] Kambhampaty, S., Chandra, S. (2006), Service oriented architecture for enterprise applications, SEPADS'06: Proceedings of the 5th WSEAS International Conference on Software Engineering, Parallel and Distributed Systems, [http://dl.acm.org/citation.cfm?id=1365739.1365748&coll=DL&dl=GUIDE&CFID=528744429&CF\\_TOKEN=40603217](http://dl.acm.org/citation.cfm?id=1365739.1365748&coll=DL&dl=GUIDE&CFID=528744429&CF_TOKEN=40603217)
- [10] Kambhampaty, S. (2007), Service oriented analysis and design process for the enterprise, ACS'07 Proceedings of the 7th Conference on 7th WSEAS International Conference on Applied Computer Science - Volume 7, [http://dl.acm.org/citation.cfm?id=1348171.1348235&coll=DL&dl=GUIDE&CFID=528744429&CF\\_TOKEN=40603217](http://dl.acm.org/citation.cfm?id=1348171.1348235&coll=DL&dl=GUIDE&CFID=528744429&CF_TOKEN=40603217)
- [11] Oliver, A. (2014), APIs will glue together the Internet of things, Javaworld, <http://www.javaworld.com/article/2606963/java-me/apis-will-glue-together-the-internet-of-things.html>

## AUTHORS

Shankar Kambhampaty is a CSC Distinguished Architect and Chief Technology Officer (CTO) for a major account at CSC, Hyderabad, India. He is author of several international conference papers including the book titled, Service Oriented Architecture for Enterprise and Cloud Applications, Wiley India publication.



Sasirekha Kambhampaty is a student of Bachelor of Technology (B.Tech.) programme in Computer Science & Engineering at Gokaraju Rangaraju Institute of Engineering & Technology (GRIET), Hyderabad, India.



*INTENTIONAL BLANK*

# COMPETENCE BUILDING FRAMEWORK REQUIREMENTS FOR INFORMATION TECHNOLOGY FOR EDUCATIONAL MANAGEMENT

Rakeshh Mohan Bhatt

Department of Computer Science and Engineering,  
HNB Garhwal University, Srinagar Uttrakhand, India  
rmbhatt77@yahoo.com

## **ABSTRACT**

*Progressive efforts have been evolving continuously for the betterment of the services of the Information Technology for Educational Management (ITEM). These services require data intensive and communication intensive applications. Due to the massive growth of information, situation becomes difficult to manage these services. Here the role of the Information and Communication Technology (ICT) infrastructure particularly data centre with communication components becomes important to facilitate these services. The present paper discusses the related issues such as competent staff, appropriate ICT infrastructure, ICT acceptance level etc. required for ITEM competence building framework considering the earlier approach for core competences for ITEM. In this connection, it is also necessary to consider the procurement of standard and appropriate ICT facilities. This will help in the integration of these facilities for the future expansion. This will also enable to create and foresee the impact of the pairing the management with information, technology, and education components individually and combined. These efforts will establish a strong coupling between the ITEM activities and resource management for effective implementation of the framework.*

## **KEYWORDS**

*ITEM Competence, Data Centre, Educational Management*

## **1. INTRODUCTION**

We have seen the strength of Information Technology (IT) in providing innovation and competitive edge to any organization (Talero & Gandette 1995). Further, using the Management Information System (MIS), many benefits have been observed in the education system (Visscher et.al. 2001), that is how the MIS systems affect the control of educational institution and educators performing administrative functions too. The use of educational information management system provides a greater degree of standardization of administrative functions. Embedded software enables a kind of automation to some extent the job to be performed without any control. But, policies of decentralization are yet to be achieved. Three levels of training (Donnelly, 2000) have been claimed as appropriate to both the leadership team and administrative

Jan Zizka et al. (Eds) : CCSIT, SIPP, AISC, CMCA, SEAS, CSITEC, DaKM, PDCTA, NeCoM - 2016

pp. 45–51, 2016. © CS & IT-CSCP 2016

DOI : 10.5121/csit.2016.60105

staff viz. training in generic software, training in the use of institution specific MIS software, and training in the use of the internet. Additionally, the leadership team should be trained to use data or information for the improvements of the educational standards. So, ICT acceptance level should be high among the members.

During the ITEM-2002, a Discussion Group was formed. This discussion group analyzed that the emerging technologies are not very well adopted by the educational institutions. These institutions are failed to provide the coherent and effective training programs as reported the unsatisfactory use of ICT means for the teaching, learning and administrative purposes (Lambert & Nolan 2002, DfES 2002, Newton 2002). Therefore, appropriate ICT infrastructure should be considered to facilitate the services required. As a result, the Discussion Group designed a model or framework to enable to plan the ITEM training and achieve an ITEM- competent staff ( Ian Selwood & et. al. 2003). The present paper analyzes the appropriate ICT infrastructure, competent staff, and ICT acceptance level required in the following sections of ICT infrastructure and ITEM competence building framework for its proper implementation.

## **2. ICT INFRASTRUCTURE**

Information management system mend for the educational institute enables central education authorities to exercise a form of “control at a distance” over the institutional operation without appearing to intervene directly. Enactment of more e-enabled services learning activities through asynchronous or synchronous processes is being effectively performed.

While analyzing the adoptability of emerging technologies, the designed framework by the Discussion Group is consisted 36 competences across the three dimensions. The first dimension points for the four inextricably linked dimensions; they are Information, Technology, Education and Management. The second dimension is set for the management and planning concerned. For this purpose, it has three levels viz. operational, tactical and strategic. The last, third dimension defines the stage of growth. Three growth labels are considered i.e. initiation, expansion and embedded. Thus, altogether this form a matrix of  $4 * 3 * 3$  and out of this, 36 competences/activities have been emerged out.

We cannot limit the workability in the collaborative environment and this phenomenon has been increased tremendously with the effective use of IT. For effective global operations, use of IT is fundamental (King & Sethi 1999) as it coordinates the dispersed activities and establish coalition in between different activities. So, organizations are trying to uncover this potential of usage of IT which could only be achieved if the organization has the appropriate technological infrastructure. For example, issues as proposed (Bhatt, 2007) such as wireless technology, cabling, Ethernet & Asynchronous Transfer Mode (ATM), iSCSI and IPv6 are important to consider to support and integrate e-communication for the massive information. In the recent past, progressive efforts have been evolving continuously for the betterment of the services of the information system for educational management, which require data intensive and communication intensive applications. Due to the massive growth of information, situation becomes difficult to manage these services and in this respect the role of ICT infrastructure particularly data centre with communication components becomes important to facilitate these services. Therefore, the emerging concept of data centre model (Clabby Analytics, 2008) can be considered to handle data intensive and data communication intensive applications becomes inevitable for the e-learning process. Till data modernization of data centre have been performed

in the business organization (Barnett, 2008). But a lot of transformation is needed in the educational institutions. This is because modern data centre are based on object-oriented technology, standard format for data sharing e.g. XML; Java technologies & internet protocol (IP) for interoperability and network communication. During the data intensive applications, large data sets are connected for distribution and may be further connected with intranet, extranet and internet. For this purpose, data centre provides data searching, retrieval and sharing quickly via its data consolidation process. For communication intensive applications, high efficient data search, retrieval and large scale collaborative multimedia system supports are required. For example, real-time multimedia interactive requires high network bandwidth for efficient data transmission in the collaborative working environment. This is achieved with the help of server virtualization process.

Therefore, through consolidation and virtualization, performance of systems and communication network for storage and sharing resources is enhanced for data as well communication intensive applications where increased traffic poses bandwidth constraints. So, to enable network responsible (McGillicuddy, 2012), server virtualization is essential to ease the operation of data centre infrastructure.

In addition to the benefits in providing standardization of administrative functions as discussed above, another kind of standardization should be considered for technological advancement so as to integrate the existing technological infrastructure components with their counterparts to come in future i.e. components should be compatible in their operations with each other. This technological integration would automatically integrate the collaborative and sharing efforts at the interaction level of the defined three axis of the model so that mapping with its existing policies and programs and investigate goodness of fit in accordance to the ITEM competencies can be achieved. In order to control the growth, the feedback process at the management level can be enforced.

### **3. COMPETENCE BUILDING FRAMEWORK FOR ITEM**

Visser and Branderhost (2001) have addressed five skill areas - to recognize the information value & policy development, to determine the type of information needed, to discover the information out of the MIS, to interpret information from the MIS, and to make decision and evaluate the policy. But, in addition to these five skills, one more skill is required. It can be named as the feedback-skill and is useful while interacting with the existing MIS at any level which would strengthen the management for its strategic, tactical and operational competences. These levels can be categorized as level-1, level-2 and level-3. The type of levels can be defined as the Level-1 is the type of structured competence; which is a prerequisite for transition to the Level-2, a semi-structured competence and Level-2 competence is a prerequisite for the transition to the Level-3, which is a kind of unstructured competence. So, the order of priority among the competences to be given is a complex task and pertains to the strategic competence; which comes at the top of all these three levels, partly can be matched with the Level-3 competence.

Further, where to initiate or expand or embed in order to retain the balance as such and for how much time. For example, strategic competence takes less time to take quick decisions; technical competence takes some more time and operate competence takes longer time consumed in providing its services and interactions processes.

Thus, to evaluate the ICT skills and pedagogy for ICT enabled teaching process, following points are needed to be analyzed:

- a) The level of acceptance of ICT skill varies from school level to post-graduate level of institutions.
- b) The level of ICT skill also reflects the level of procurement and integration of ICT equipments.
- c) ICT skill is also influenced by the availability of the ICT facilities exist in the institute.

These points should be considered the most important activity, hence the management team/authority (comprising educationists and technocrats) should have a very broad vision so as to open ICT windows at all levels of learning and teaching to enhance the availability of the ICT usage. This phenomenon will also empower the management process through the feed back received instantly by building the ICT means besides the ICT skill. However, this empowerment will also be benefited by the advantages from the proposed model (Ian Selwood et.al., 2003) such as it is platform independent, and it is descriptive and prescriptive. In meeting the three-group competencies such as using tools interactively, interacting in heterogeneous groups and acting autonomously, teaching process should also be under the adequate educational environment (Kollee, C. et. al., 2009). Therefore, suitable mapping can be done with the existing activities for its fitness.

The above analysis covering all the three points would help the strategic support to distribute the ICT resources uniformly. Essentially a problem of management, relates with the process of managing the information systems can be implemented by providing best support technologies components to maintain the richness of the information. So, this will ease the management work in the proposed model of ITEM competence by considering the relevance and compatible technology for education and vis-à-vis segregate them ( information, technology and education) for strategic, tactical and operational levels of actions. Then the steps towards the initiation, expansion and embedment can be taken accordingly.

This would lead to the ITEM competence building properly. Therefore, it is recommended for the management process to analyze the levels of:

- a) acceptance of ICT skill,
- b) procurement and integration of ICT tools/ equipments, and
- c) ICT facilities,

This analysis will help to create and foresee the impact of the pairing the management with information, technology, and education components individually and combined. This would also improve the base for the ITEM training and achieve an ITEM-competent technology and staff. This will then become easy to anticipate the inevitable changes. Attributed benefits accruing out of it can also be traced and understood.

To extract standards to support teaching and learning for ITEM “Technological Standards for School Administrators”(TSSA collaborative 2001) is suitable approach. Further, Bhatt (2007) has pointed out that the ITEM-competence is also strongly influenced by the appropriate selection and use of information and technological components. Therefore, this kind of integration, data centre concept and the TSSA approach can be considered for the better ITEM competence building. Further, the UNESCO ICT Competency Framework (Wallet, 2014) has also been designed for Teachers which is a useful tool to inform education policymakers, educators and providers of professional learning of the role of ICT in educational reform. It also assists Member States in developing national ICT competency standards. Emphasis has been given on to collaborate in problem-solving and creative learning so as to enhance the student outcomes. In this connection, it is also necessary to consider the procurement of standard and appropriate ICT facilities. This will help in the integration of these facilities for the future expansion. This will also enable to create and foresee the impact of the pairing the management with information, technology, and education components individually and combined. These efforts will establish a strong coupling between the ITEM activities and resource management for effective implementation of the framework.

#### **4. CONCLUSIONS**

Legacy storage and communication support systems should be transformed to modernize the data centre in order to resolve constantly changing demands for various kinds of applications overcoming the interoperability, data sharing problems and quality of information delivery. The aforesaid efforts could provide the benefits in the proposed structured approach of the model by solving the complex, dynamic and distributed behavior of the information operations with the help of virtualization, data consolidation and network management. Therefore, for the better competences and to accelerate the interactions more effectively, it is strongly believed that with appropriate implementation of technological components and analyses of management with information, technology, and education components individually and combined; a strong coupling can be established between the ITEM activities and resource management to nurture the ITEM building process.

#### **ACKNOWLEDGEMENTS**

The author is sincerely thankful to HNB Garhwal University, Srinagar and Utrakhand State Council for Science & Technology for providing financial support to present this paper.

#### **REFERENCES**

- [1] Barnett, Jef, (2008) “The New Enterprise Data Centre: Enabling Business Innovation”, IBM Corporation, USA.
- [2] Bhatt, R. M. (2007), “Communication Support Technologies for e-Learners”, Knowledge Management for Education Management, eds. Tatnall, A, Okamoto, T., Visscher, A., (Boston:Springer), 230: 69-73.
- [3] Clabby Analytics (2008), “The Data Centre ‘Implosion Explosion’... and the Need to Move to a New Enterprise Data Centre Model”, Feb., 2008 (<http://clabbyanalytics.com>) (Boston:Springer), 230: 69-73.

- [4] DfES (2002a). “Staff ICT Competences framework”, 21 Aug.  
<http://www.teachernet.gov.uk/docbank/index.cfm?id=2820>. DfES. London
- [5] DfES (2002b). “Who does what”, 21 Aug.  
<http://www.teachernet.gov.uk/docbank/index.cfm?id=28201>. DfES. London
- [6] Donnelly, J. (2000). “Information Management Strategy for Schools and Local Education Authorities – Report on Training Needs”, <http://dfes.gov.uk/ims/JDReportfinal.rtf>. DfES, London.
- [7] Ian Selwood and O’Mahony, D. Christopher with Rakesh Bhatt, Bill Davy, Margaret, Hatano Kazuhiko, Javier Osorio and Tuulikki Paturi (2003). “ Management of Education in the Information Age : The Role of ICT”, eds. Ian D. Selwood, Alex C.W. Fung, Christopher D. O’Mahony, Kluwer for IFIP. USA, 195-201.
- [8] King W.R. and Sethi V. (1999). “An empirical assessment of the organization of transnational information system”. *Journal of Management Information System* 15(4), 7-28.
- [9] Kollee, C.,Magenheim, J.,Nelles, W., Rhode, T., Schaper, N., Schubert, Sirgid, and Stechert Peer, (2009). “Computer Science Education and Key Competencies”, in *Proced of IFIP WCCE 2009*,147-156.
- [10] Lambert M.J. and Nolan, C.J.P. (2003). “Managing learning environment in schools”. *Management of Education in The Information Age – The Role of ICT*, Ed. Selwood I, Fung A, O’Mahony C. Kluwer for IFIP, London.
- [11] Newton L. (2003). “Management and the use of ICT in subject teaching integration for learning”. *Management of Education in The Information Age – The Role of ICT*, Ed. Selwood I, Fung A, O’Mahony C. Kluwer for IFIP, London.
- [12] McGillicuddy, S. (2012), “Data Centre Network Fabrics vs. Software-Defined Network“, *Network Evolution E-Zine*, 3(4),pp. 6-10.
- [13] O’Mahony, D. Christopher (2007). “Knowledge Management for Education Management”, eds. Tatnall, A, Okamoto, T., Visscher, A., (Boston:Springer), 230: 69-73.
- [14] Talero & Gandette (1995), “Harnessing information for development: A proposal for World Bank Group Vision and Strategy”, *IT for Development*, 6:145-188.
- [15] TSSA (2001). “Technological Standards for School Administrators”.  
<http://cnets.iste.org/tssa/docs/tssa.pdf>.TSSA Collaborative/ISTE, Engene.
- [16] Visscher, A, J. and Brandhost, E.M. (2001). “How should school managers be trained for managerial school information system usage?” , In *Pathways to Institutional Improvement with Information Technology in Educational Management*. Eds. Nolan C.J.P., Fung,A.C.W., & Brown,M.A. Kluwer for IFIP.London.
- [17] Wallet, Peter, (2014). “Information and Communication Technology in Asia : A comparative analysis of ICT integration and e-readiness in schools across Asia”, Published by: UNESCO Institute for Statistics, P.O. Box 6128, Succursale Centre-Ville Montreal, Quebec H3C 3J7 Canada, Ref: UIS/2014/ICT/TD/3 REV , DOI <http://dx.doi.org/10.15220/978-92-9189-148-1-en>

## **AUTHORS**

### **Dr Rakesh Mohan Bhatt**

Dr R M Bhatt has experience of about 30 years in the field of Computer Applications. He is associated with the IFIP WG on AI, and Education. He is also a senior life member of Computer Society of India. He has been Visiting Professor at SS Cyril University, Slovakia and recipient of Commonwealth Fellowship Award for Academic Exchange. He has published around 45 papers.



*INTENTIONAL BLANK*

# STABLE MARRIAGE PROBLEM WITH TIES AND INCOMPLETE BOUNDED LENGTH PREFERENCE LIST UNDER SOCIAL STABILITY

Ashish Shrivastava and C. Pandu Rangan

Department of Computer Science and Engineering,  
Indian Institute of Technology, Madras, Chennai, India  
ashish3586@gmail.com, prangan55@gmail.com

## ABSTRACT

*We consider a variant of socially stable marriage problem where preference lists may be incomplete, may contain ties and may have bounded length. In real world application like NRMP and Scottish medical matching scheme such restrictions arise very frequently where set of agents (man/woman) is very large and providing a complete and strict order preference list is practically in-feasible. In presence of ties in preference lists, the most common solution is weakly socially stable matching. It is a fact that in an instance, weakly stable matching can have different sizes. This motivates the problem of finding a maximum cardinality weakly socially stable matching.*

*In this paper, we find maximum size weakly socially stable matching for an instance of Stable Marriage problem with Ties and Incomplete bounded length preference list with Social Stability. The motivation to consider this instance is the known fact, any larger instance of this problem is NP-hard.*

## KEYWORDS

*Stable Marriage Problem, Socially Stable Matching, Bipartite Matching, Stable Marriage Problem with Ties and Incomplete list.*

## 1. INTRODUCTION

The *Stable marriage problem* was first introduced by Gale and Shapley in 1962 [1]. The *classical* instance  $I$  of the stable marriage problem has a set of  $n$  men  $U$ , a set of  $n$  women  $W$  and *preference lists* of men over women and vice versa. Each preference list contains all members of opposite sex in a strict order. A man  $m_i$  and a woman  $w_j$  are called *acceptable* to each other in instance  $I$  if  $m_i$  is in preference list of  $w_j$  and  $w_j$  is in preference list of  $m_i$ . Let  $\alpha$  is the set of all *acceptable pairs* in the instance  $I$ . A *matching*  $M$  is a set of independent pairs  $(m_i, w_j)$  such that  $m_i \in U$  and  $w_j \in W$ . If  $(m_i, w_j) \in M$ , we say that  $m_i$  is matched to  $w_j$  in  $M$  and vice versa and we denote  $M(m_i) = w_j$  and  $M(w_j) = m_i$ .

A pair  $(m_i, w_j) \notin M$  is called a *blocking pair* for matching  $M$  if both  $m_i$  and  $w_j$  prefer each other to their partners in  $M$ . A matching  $M$  is called a *stable matching* iff there is no blocking pair with respect to  $M$ . Gale and Shapley gave a deferred acceptance algorithm and proved that every instance  $I$  of the stable marriage problem admits a stable matching which can be found in polynomial time [1].

The largest and one of the best known applications of Hospitals Residents problem is National Resident Matching Program (NRMP) and Scottish medical matching scheme which match graduated medical students (residents) with their preferred hospitals on the basis of both side preference lists.

The research work in the field of The Stable Marriage Problem has a long history. As we have mentioned earlier, the first problem on stable marriage problem was introduced by Gale and Shapley in 1962. After that lots of variation on first problem came into the picture. Some major variations are Stable Marriage problem with Ties (SMT), Stable Marriage problem with Incomplete list (SMI), Stable Marriage problem with Ties and Incomplete list (SMTI) and Stable Marriage problem with Bounded length preference lists.

### 1.1 Stable marriage problem with ties (SMT)

In Stable Marriage problem with Ties, each man can give a preference list over a set of women, where two or more women can hold the same place (*ties*) in the preference list and vice-versa. In SMT there are three notion of stability: weak stability, strong stability and super stability [2, 3]. A blocking pair  $(m_i, w_j) \notin M$  with respect to a *weakly stable matching*  $M$  can be defined as follows: (a)  $m_i$  and  $w_j$  are acceptable to each other. (b)  $m_i$  strictly prefers  $w_j$  to  $M(m_i)$  (partner of  $m_i$  in matching  $M$ ) (c)  $w_j$  strictly prefers  $m_i$  to  $M(w_j)$ . For an instance  $I$  of weakly stable matching problem, a weakly stable matching  $M$  always exist and can be found in polynomial time [3].

A blocking pair  $(m_i, w_j) \notin M$  with respect to a *strongly stable matching*  $M$  can be defined as follows: (a)  $m_i$  and  $w_j$  are acceptable to each other. (b)  $m_i$  strictly prefer  $w_j$  to  $M(m_i)$  and  $w_j$  is indifferent between  $m_i$  and  $M(w_j)$  and vice-versa.

A blocking pair  $(m_i, w_j) \notin M$  with respect to a *super stable matching*  $M$  can be defined as follows: (a)  $m_i$  and  $w_j$  are acceptable to each other. (b) both  $m_i$  and  $w_j$  either strictly prefer each other to their partners  $M$  or indifferent between them. There could be an instance  $I$  that have neither super nor strongly stable matching but there is an algorithm which can find super and strong stable matching in  $I$  (if exist) in polynomial time [4]. Among these three stability notions, *weak stability* has received most attention in the literature [5-12].

### 1.2 Stable marriage problem with incomplete lists (SMI)

Stable Marriage with Incomplete list (SMI) is another variation of stable marriage problem in which number of men and women in an instance  $I$  need not be same. Each man and woman can give a preference list over a subset of opposite sex. For an instance  $I$  a pair  $(m_i, w_j)$  is called blocking pair with respect to a matching  $M$  if: (a)  $m_i$  and  $w_j$  are acceptable to each other (b)  $m_i$  is either unmatched in  $M$  or prefer  $w_j$  to  $M(m_i)$  (c)  $w_j$  is either unmatched in  $M$  or prefer  $m_i$  to  $M(w_j)$ . A matching  $M$  is called stable if there is no blocking pair with respect to  $M$ .

In an instance  $I$  of SMI we can partition the set of men and women such that, one partition have those men and women which have partners in all stable matching and other partition have those

men and women which are unmatched in all stable matching [13].

### 1.3 Stable marriage problem with ties and incomplete lists (SMTI)

Stable Marriage with Ties and Incomplete list (SMTI) is an extension of classical stable marriage problem in which number of men and women in an instance  $I$  need not be same. Each man gives a preference list over a subset of women and vice-versa. Each preference list may contain ties (two or more men/women have same rank). A pair  $(m_i, w_j) \notin M$  forms a blocking pair with respect to matching  $M$  if (a) Both  $m_i$  and  $w_j$  are acceptable to each other and (b)  $m_i$  is either unmatched or strictly prefers  $w_j$  to  $M(m_i)$  and (c)  $w_j$  is either unmatched or strictly prefers  $m_i$  to  $M(w_j)$ . A matching  $M$  is called a weakly stable matching if there is no blocking pair with respect to  $M$ .

It is known that a weakly stable matching in an instance  $I$  of SMTI can have different sizes and finding maximum cardinality weakly stable matching is an NP-hard problem [6]. NP-hardness holds even if only one tie of size 2 occurs on men's preference list at the tail and women's preference list contain no ties [6].

### 1.4 Stable marriage problem with bounded length preference lists.

The idea behind bounded length preference list is, in case of large scale matching problems, the preference list of at-least one side of agent tend to be short. An example of large scale matching is Scottish medical matching scheme [14] where each student is required to rank only three hospitals in their preference list. This variation leads to a question, whether problem of finding maximum size stable matching becomes simpler? (For an instance, with one side or both sided bounded preference list).

Suppose  $(p, q)$ -MAX SMTI denotes such variation on MAX SMTI problem (finding maximum size matching in an instance of SMTI) where each man can give at-most  $p$  women in his preference list and each woman can give at-most  $q$  men in her preference list. Halldorsson et al. [7] showed that  $(4, 7)$ -MAX SMTI is NP-hard and not approximable within some  $\delta > 1$  unless  $P = NP$ . A reduction from Minimum Vertex Cover to MAX SMTI, shows that later problem cannot be approximable within  $21/19$  unless  $P = NP$  [9]. Another study in [15] uses NP-hard restriction of minimum vertex cover of graph of minimum degree 3 in producing NP-hard result for  $(5, 5)$ -MAX SMTI. Irving et al. [16] shows that  $(3, 4)$ -MAX SMTI is NP-hard and not approximable within  $\delta > 1$  unless  $P = NP$ .

## 2. RELATED WORK

Another variation of stable marriage problem is socially stable marriage problem. An instance  $I'$  of socially stable marriage problem can be defined by  $(I, G)$  where  $I$  is an instance of classical stable marriage problem and  $G = (U \cup W, A)$  is a social network graph. Here  $U$  and  $W$  are set of men and women respectively and  $A$  is set of man woman pair who knows each other in social network  $G$ . Set  $A$  is called the set of *acquainted pairs* which is the subset of all acceptable pairs ( $A \subseteq \alpha$ ). A marriage  $M$  is called socially stable marriage if there is no socially blocking pair with respect to  $M$ . A socially blocking pair  $(m_i, w_j) \notin M$  is defined as follows: (a) both  $m_i$  and  $w_j$  prefers each other to their partner in  $M$  and (b)  $m_i$  and  $w_j$  are connected in social network  $G$ .

In large scale matching like NRMP and Scottish medical matching scheme, social stability is a useful notion in which members of blocking pair block a matching  $M$  only if they know the

existence of each other. Thus the notion of social stability allows us to increase the cardinality of matching without taking care of those pairs which are not socially connected in social network graph.

The work in this paper is motivated by the work of Irving et al. [16] where they study about stable marriage problem with ties and bounded length preference list. They show that if each man's list is of length at most two and women's lists are of unbounded length with ties, we can find a maximum size weakly stable matching in polynomial time.

Our work in this paper is also motivated by the work of Askaladis et al. [17] where they study about socially stable matching problem with bounded length preference list. They gave a  $O(n^{3/2} \log n)$  time algorithm for  $(2, \infty)$ -MAX SMISS problem. Where  $(2, \infty)$ -MAX SMISS problem is to find a maximum size socially stable matching in an instance of stable marriage problem with incomplete list under social stability, where each man's list is of length at most two (without ties) and women's lists are of unbounded length (without ties).

### 3. OUR CONTRIBUTION

In an instance  $I$  of  $(2, \infty)$ -MAX SMISS problem if we include ties on both side preference lists, where the length of a tie could be arbitrary, this instance converts into an instance  $I'$  of  $(2, \infty)$ -MAX SMTISS. In this paper we will show that we can find maximum size weakly socially stable matching in instance  $I'$  in polynomial time. Due to presence of ties in both side preference lists there are three notion of stability: weak, strong and super. In this paper we are considering maximum size weakly stable matching in  $I'$  of  $(2, \infty)$ -MAX SMTISS.

As we mention earlier, Irving et al. [16] shows that  $(3, 4)$ -MAX SMTI is NP-hard and not approximable within  $\delta > 1$  unless  $P = NP$ . It follows that the complexity status of  $(3, \infty)$ -MAX SMTI is also NP-hard. Similarly socially stable variation of  $(3, \infty)$ -MAX SMTI problem, “ $(3, \infty)$ -MAX Weakly SMTISS” is also NP-hard.

Given an instance  $I'$  of  $(2, \infty)$ -MAX Weakly SMTISS (Stable Marriage problem with Ties and Incomplete bounded length preference list under Social Stability), we present an algorithm that gives a maximum size weakly socially stable matching with time complexity  $O(n^{3/2} \log n)$ , where  $n$  is the total number of men and women in the instance  $I$ .

### 4. STABLE MARRIAGE PROBLEM WITH TIES AND INCOMPLETE BOUNDED LIST UNDER SOCIAL STABILITY (SMTISS)

An instance of Stable Marriage Problem with Ties and Incomplete bounded list under Social Stability (SMTISS) can be defined by  $(I, G)$  where  $I$  is the instance of SMTI and  $G = (U \cup W, A)$ , where  $A$  (the set of all acceptable pairs). A man  $m_i$  and a woman  $w_j$  are called *socially connected* to each other in graph  $G$  if  $(m_i, w_j) \in A$ . Each preference list is a partial order on a subset of opposite sex. A matching  $M$  is called *weakly socially stable* if there is no socially blocking pair. A pair  $(m_i, w_j) \notin M$  is a socially blocking pair if (a)  $(m_i, w_j) \in A$  and (b)  $m_i$  is either unmatched or strictly prefers  $w_j$  to his partner in  $M$  and (c)  $w_j$  is either unmatched or strictly prefers  $m_i$  to her partner in  $M$ . In general, for any instance  $I$  of SMTISS problem, one of the aim is to compute a maximum cardinality socially stable matching (weakly, strong, super etc). In an incomplete tied preference list, arbitrary breaking of ties need not always lead to a maximum weakly socially stable matching. The following example shows that if we break ties arbitrarily we can find

weakly socially stable matching of different sizes.

<b>Example:</b>	Men's preference lists	Women's preference lists
	$m_1: (\underline{w_1}, w_2)$	$w_1: \underline{m_1}, m_2$
	$m_2: w_1$	$w_2: m_1$

In above example the underline shows a social connection in  $G$ . Here man  $m_1$  has a social connection with woman  $w_1$ . Observe that if we break the tie of  $m_1$  as  $m_1 : \underline{w_1}, w_2$  then maximum weakly socially stable matching will be  $\{(m_1, w_1)\}$  of size 1 and if we break tie of  $m_1$  as  $m_1 : w_2, \underline{w_1}$  then maximum weakly socially stable matching will be  $\{(m_1, w_2), (m_2, w_1)\}$  of size 2. The above example motivates us to find maximum cardinality weakly socially stable matching in an instance  $I$  of SMTISS.

Observe that if we restrict the length of all ties equal to 1 in an instance  $I$  of SMTISS then it will reduce into an instance  $I'$  of SMISS. Since it is known that finding a maximum cardinality socially stable matching in an instance of SMISS is NP-complete [17], finding a maximum cardinality weakly socially stable matching in an instance of SMTISS is also NP-complete. Askalidis et al. showed that the problem  $(2, \infty)$ -MAX SMISS ( $(2, \infty)$ -MAX SMTISS with ties length 1) is solvable in polynomial time [17], this result directed us to a more general version called  $(2, \infty)$ -MAX Weakly SMTISS problem where ties length could be two or more. It may seem that one can consider that if we break the ties arbitrary and apply  $(2, \infty)$ -Max SMISS algorithm then we can find maximum cardinality weakly socially stable matching for  $(2, \infty)$ -SMTISS instance, but this is not always true. We can verify this by above example.

#### 4.1 ALGORITHM FOR $(2, \infty)$ -MAX WEAKLY SMTISS

The objective of this problem is to find a maximum cardinality weakly socially stable matching in SMTI instance under social stability, where each man can give a preference list of length at most two and each woman can give unbounded length incomplete list, with or without ties of any length. We present an  $O(n^{3/2} \log n)$  time algorithm for this problem. Similar to  $(2, \infty)$ -MAX SMISS given in [17], this algorithm also completes in three phases. In phase 1 we delete all pairs which can never belong to any weakly socially stable matching. The intuition behind phase 1 is, if there is a man  $m_i$  who is socially connected to his first choice woman  $w_j$  then any man who is less preferable than  $m_i$  in  $w_j$  preference list, cannot match with  $w_j$  in any socially stable matching. If it happens,  $(m_i, w_j)$  will be blocking pair for resultant socially stable matching  $M$ .

In phase 2, first we build a graph from the reduced instance from phase 1 and weight each edge  $(m_i, w_j)$  by  $rank(w_j, m_i)$ , where  $rank(w_j, m_i)$  is rank of man  $m_i$  in  $w_j$ 's reduced preference list. Now we construct a minimum weight maximum matching  $M_G$  in graph. Finally, in phase 3 we settle those pairs which are matched in phase 2 but will be socially blocking pair for output matching  $M$ .

**Lemma 4.1.1.**  $(2, \infty)$ -MAX weakly SMTISS algorithm terminates.

**Proof.** We start phase 1 by unmarking all men. Now we mark those men who are unmarked and have a non-empty reduced list. When every man becomes either marked or having an empty reduced preference list, phase 1 will terminate. Since a man  $m_i$  can be marked at most twice

during phase 1 and total number of men in instance  $(I, G)$  is finite, phase 1 will terminate. In phase 2 of algorithm we are finding a minimum weight maximum matching of the reduced instance, therefore phase 2 will also terminate. In phase 3, each iteration improves the choice of a man from his second choice woman to his first choice woman and no man obtains worse woman or becomes unmatched. Since total number of possible improvements for men is finite, therefore the total number of iterations is also finite and hence phase 3 will also terminate.

<p style="text-align: center;">/ Phase 1 /</p> <p>Set all men to be unmarked;</p> <p><b>while</b> some man <math>m_i</math> is unmarked and <math>m_i</math> has a non-empty reduced list <b>do</b></p> <p style="padding-left: 20px;">Set <math>m_i</math> to be <b>marked</b>;</p> <p style="padding-left: 20px;"><b>if</b> <math>m_i</math>'s reduced list is not a tie of length 2 <b>then</b></p> <p style="padding-left: 40px;"><math>w_j :=</math> woman in first position on <math>m_i</math>'s reduced list;</p> <p style="padding-left: 40px;"><b>if</b> <math>(m_i, w_j) \in A</math> <b>then</b></p> <p style="padding-left: 60px;"><b>for</b> each successor <math>m_k</math> of <math>m_i</math> on <math>w_j</math>'s list <b>do</b></p> <p style="padding-left: 80px;">Set <math>m_k</math> to be <b>unmarked</b> ;</p> <p style="padding-left: 80px;">Delete pair <math>(m_k, w_j)</math>;</p>	<p style="text-align: center;">/ Phase 3 /</p> <p><math>M = M_G</math>;</p> <p><b>while</b> (there exists a man <math>m_i</math> who is assigned to his second choice woman <math>w_k</math> and his first choice woman <math>w_j</math> is free in <math>M</math> and <math>(m_i, w_j) \in A</math>) <b>do</b></p> <p style="padding-left: 20px;"><math>M = M \setminus \{(m_i, w_k)\}</math>;</p> <p style="padding-left: 20px;"><math>M = M \cup \{(m_i, w_j)\}</math>;</p> <p>Return <math>M</math>;</p> <p><b>Build Graph</b>();</p> <p><math>V = U \cup W</math>;</p> <p><math>E = \phi</math>;</p> <p><b>for</b> each man <math>m_i \in U</math> <b>do</b></p> <p style="padding-left: 20px;"><b>for</b> each woman <math>w_j</math> on <math>m_i</math> reduced list <b>do</b></p> <p style="padding-left: 40px;"><math>E = E \cup (m_i, w_j)</math>;</p> <p style="padding-left: 40px;"><math>weight(m_i, w_j) = rank(w_j, m_i)</math>;</p> <p><math>G = (V, E)</math>;</p> <p>Return <math>G</math>;</p>
<p style="text-align: center;">/ Phase 2 /</p> <p><math>G = \text{Build Graph}()</math>;</p> <p><math>M_G =</math> Minimum weight maximum matching in <math>G</math>;</p>	

Figure 1.  $(2, \infty)$ -MAX Weakly SMTISS Algorithm

**Lemma 4.1.2.** Phase 1 of  $(2, \infty)$ -MAX Weakly SMTISS Algorithm never deletes a weakly socially stable pair.

**Proof.** Suppose  $(m_i, w_j)$  is a weakly socially stable pair which has been deleted during execution of phase 1 of algorithm 1 such that  $(m_i, w_j) \in M$ , where  $M$  is a weakly socially stable matching in  $(I, G)$ . Suppose this is the first weakly stable pair deleted during phase 1. This deletion was done because of  $w_j$  being the first choice of some man  $m_r$  where  $(m_r, w_j) \in A$ ,  $m_r$ 's reduced list was not a tie of length 2 and  $w_j$  prefers  $m_r$  to  $m_i$ . But in that case pair  $(m_r, w_j)$  becomes a social blocking pair with respect to matching  $M$ . This is a contradiction to the fact that  $M$  is a weakly socially stable matching.

**Lemma 4.1.3.** The matching returned by algorithm  $(2, \infty)$ -MAX weakly SMTISS is weakly socially stable in  $(I, G)$ .

**Proof.** Suppose our algorithm outputs the matching  $M$  and this matching is not weakly socially stable in  $(I, G)$ . It means there exist a pair  $(m_i, w_j)$  which is a socially blocking pair with respect to  $M$ . We can consider following four cases corresponding to a socially blocking pair.

**Case (i)** both  $m_i$  and  $w_j$  are unmatched in  $M$ :

We know once a man  $m_i$  is matched in  $M_G$ , he will never be unmatched in phase 3. Either  $m_i$  remains with his partner in  $M_G$  or he can improve his partner (if possible) during phase 3. Therefore, if a man  $m_i$  is unmatched in  $M$ , he was unmatched in  $M_G$ . Woman  $w_j$  can either be unmatched in  $M_G$  or during phase 3. In first case, suppose a woman  $w_j$  is unmatched in  $M_G$ , then we can increase the size of matching  $M_G$  by adding the edge  $(m_i, w_j)$ , which contradicts the fact that matching  $M_G$  is a maximum matching. In second case, suppose a woman  $w_j$  becomes unmatched during phase 3, then it means that her partner in  $M_G$ , say  $m_{p1}$ , had a strict preference list of length 2 and got his first choice woman, say  $w_{q1}$ , where  $(m_{p1}, w_{q1}) \in A$ . Now again we have two cases for  $w_{q1}$ . In the first case, woman  $w_{q1}$  is unmatched in  $M_G$ . This leads to an augmenting path  $\{(m_i, w_j), (m_{p1}, w_j), (m_{p1}, w_{q1})\}$  in  $M_G$ , where the first and the last edges are not in  $M_G$ . So we can increase the size of the matching  $M_G$  by one. This is a contradiction to the fact that  $M_G$  is maximum matching. In second case, suppose  $w_{q1}$  becomes unmatched during phase 3, then it means her partner in  $M_G$ , say  $m_{p2}$ , had a strict preference list of size 2 and got his first choice woman  $w_{q2}$ , where  $m_{p2}$  and  $w_{q2}$  had a social connection in  $(I, G)$ . Again we can observe an augmenting path  $\{(m_i, w_j), (m_{p1}, w_j), (m_{p1}, w_{q1}), (m_{p2}, w_{q1}), (m_{p2}, w_{q2})\}$  in  $M_G$  which contradicts the fact that  $M_G$  is maximum matching in  $(I, G)$ . Similarly, if we keep on doing this operation, number of men is finite and since every man strictly improves his partner in  $M_G$ , there exist a finite number of women who can become unmatched after phase 3. Hence at some time there exists a man  $m_{pr}$ , who is matched with  $w_{q_{r-1}}$  in  $M_G$ , and switched to his first choice  $w_{qr}$  and  $w_{qr}$  is unmatched in  $M_G$ . Here we can form an augmenting path  $\{(m_i, w_j), (m_{p1}, w_j), (m_{p1}, w_{q1}), (m_{p2}, w_{q1}), (m_{p2}, w_{q2}), \dots, (m_{pr}, w_{qr})\}$ , which leads to a contradiction that  $M_G$  is a maximum matching.

**Case (ii)**  $m_i$  is unmatched in  $M$  and  $w_j$  prefers  $m_i$  to  $M(w_j)$ :

As explained before  $m_i$  is unmatched in  $M_G$ . Woman  $w_j$  is either matched to  $M(w_j)$  in  $M_G$  or matched to some  $m_{p1}$  in  $M_G$  and after that matched to  $M(w_j)$  in phase 3. In first case, if  $w_j$  is matched to  $M(w_j)$  in  $M_G$  then it leads to contradiction that  $M_G$  is a minimum weight maximum matching. We can simply discard the edge  $(M(w_j), w_j)$  from  $M_G$  and add edge  $(m_i, w_j)$ , which reduces the weight of  $M_G$  without reducing its cardinality. In second case,  $w_j$  is matched to some  $m_{p1}$  in  $M_G$ , where  $m_{p1} \neq m_i \neq M(w_j)$  and after that matched to  $M(w_j)$  in phase 3. Now, after phase 3,  $w_j$  is not matched with  $m_{p1}$ , which means that  $m_{p1}$  got his first choice woman say  $w_{q1}$  in phase 3. Now woman  $w_{q1}$  is either free or matched to some man in  $M_G$ . In both cases, using similar arguments as in Case (i) we can construct an augmenting path and contradict that  $M_G$  is a maximum matching.

**Case (iii)**  $m_i$  is matched to  $M(m_i)$  in  $M$ ,  $m_i$  prefers  $w_j$  to  $M(m_i)$  and  $w_j$  is unmatched in  $M$ :

We know that man  $m_i$  has a strict preference list of size 2. It follows that  $w_j$  is the first woman of  $m_i$ 's preference list. Since  $(m_i, w_j)$  is an edge in social graph  $G$ , this satisfies the loop condition of phase 3 and  $m_i$  will be matched to  $w_j$  during phase 3. Therefore this case will never occur after execution of this algorithm.

**Case (iv)**  $m_i$  is matched with  $w_k = M(m_i)$ , and  $m_i$  prefers  $w_j$  to  $w_k$  and  $w_j$  is assigned to  $m_r = M(w_j)$  and  $w_j$  prefers  $m_i$  to  $m_r$ :

We know that length of preference list of  $m_i$  is 2 and  $m_i$  is in social connection with  $w_j$ .  $m_i$  strictly prefers  $w_j$  to  $w_k$ , which means that  $w_j$  is first choice of  $m_i$ . Woman  $w_j$  strictly prefers  $m_i$  to  $m_r$ . Therefore the loop condition of phase 1 will be true and phase 1 will delete the pair  $(m_r, w_j)$ . Hence this case will never occur in our algorithm.

Since by lemma 1 we know that phase 1 of algorithm never deletes a socially stable pair, in phase 2 we constructed a minimum weight maximum matching  $M_G$  from the reduced preference list by phase 1 using algorithm in [18]. During phase 3 we never decrease the size of  $M_G$ , which follows that resultant matching  $M$  after phase 3 is a maximum cardinality matching. Lemma 3 ensures that the matching produced by  $(2, \infty)$ -MAX weakly SMTISS algorithm is weakly socially stable. It follows that the algorithm produced a maximum weakly socially stable matching in instance  $(I, G)$ . The running time complexity of the algorithm is dominated by phase 2 which constructs a minimum weight maximum matching in  $G'(V, E')$  in time  $O(\sqrt{|V|} |E'| \log |V|)$  [18]. Suppose  $|V| = n = n_1 + n_2$  is total number of men and women, then the cardinality of set of acceptable pairs is at most  $2n_1 = O(n)$ . It follows that the time complexity of  $(2, \infty)$ -MAX weakly SMTISS algorithm is  $O(n^{3/2} \log n)$ .

**Theorem 4.1.** For a given instance  $(I, G)$  of  $(2, \infty)$ -MAX Weakly SMTISS, Algorithm  $(2, \infty)$ -MAX weakly SMTISS produces a maximum size weakly socially stable matching in  $O(n^{3/2} \log n)$  time, where  $n$  is the total number of men and women in  $I$ .

## 5. CONCLUSION AND FUTURE WORK

In this paper we have presented an algorithm for an instance of stable marriage problem with ties and incomplete bounded length preference list, where each man can give at most 2 women in his preference list (with or without ties) and each woman can give unbounded length preference list (with or without ties). Length of ties in women preference list could be 2 or more. We have found that this instance can be solved in polynomial time. These instances are very common in real world scenario like NRMP and Scottish medical matching scheme where medical students can give small size preference list. It would be interesting to study about maximum size strongly stable matching and super stable matching in the scenario of social stability. We leave this as an open problem.

## ACKNOWLEDGEMENTS

The authors are grateful to Dr. Meghana Nasre for useful discussion on the problem and anonymous reviewers for their useful comments which improved the quality and presentation considerably.

## REFERENCES

- [1] Gale, D., Shapley, L.S.: College admissions and the stability of marriage. The American Mathematical Monthly 69 (1962) 9-15
- [2] Gusfield, D., Irving, R.W.: The Stable Marriage Problem: Structure and Algorithms. MIT Press, Cambridge, MA, USA (1989)

- [3] Irving, R.W.: Stable marriage and indifference. In: Selected Papers of the Conference on Combinatorial Optimization. CO89, Amsterdam, The Netherlands, The Netherlands, North-Holland Publishing Co. (1994) 261-272
- [4] Kavitha, T., Mehlhorn, K., Michail, D., Paluch, K.E.: Strongly stable matchings in time  $o(nm)$  and extension to the hospitals-residents problem. *ACM Trans. Algorithms* 3 (2007)
- [5] Iwama, K., Manlove, D., Miyazaki, S., Morita, Y.: Stable marriage with incomplete lists and ties. In: In Proceedings of ICALP 99: the 26th International Colloquium on Automata, Languages and Programming, Springer-Verlag (1999) 443-452
- [6] Manlove, D.F., Irving, R.W., Iwama, K., Miyazaki, S., Morita, Y.: Hard variants of stable marriage. *Theoretical Computer Science* 276 (2002) 261-279
- [7] Halldrsson, M.M., Irving, R.W., Iwama, K., Manlove, D.F., Miyazaki, S., Morita, Y., Scott, S.: Approximability results for stable marriage problems with ties. *Theoretical Computer Science* 306 (2003) 431-447
- [8] Halldrsson, M.M., Iwama, K., Miyazaki, S., Morita, Y.: Inapproximability results on stable marriage problems. In Rajsbaum, S., ed.: *LATIN*. Volume 2286 of *Lecture Notes in Computer Science*., Springer (2002) 554-568
- [9] Halldrsson, M.M., Iwama, K., Miyazaki, S., Yanagisawa, H.: Improved approximation results for the stable marriage problem. *ACM Trans. Algorithms* 3 (2007)
- [10] Halldrsson, M.M., Iwama, K., Miyazaki, S., Yanagisawa, H.: Randomized approximation of the stable marriage problem. *Theoretical Computer Science* 325 (2004) 439-465 Selected Papers from {COCOON} 2003.
- [11] Iwama, K., Miyazaki, S., Okamoto, K.: A  $(2-c(\log n/n))$ -approximation algorithm for the stable marriage problem. In Hagerup, T., Katajainen, J., eds.: *SWAT*. Volume 3111 of *Lecture Notes in Computer Science*., Springer (2004) 349-361
- [12] Iwama, K., Miyazaki, S., Yamauchi, N.: A  $(2-c(1/\sqrt{n}))$ -approximation algorithm for the stable marriage problem. *Algorithmica* 51 (2008) 342-356
- [13] Gale, D., Sotomayor, M.: Some remarks on the stable matching problem. *Discrete Applied Mathematics* 1 (1985) 223-232
- [14] Irving, R.W.: Matching medical students to pairs of hospitals: A new variation on a well-known theme. In Bilardi, G., Italiano, G.F., Pietracaprina, A., Pucci, G., eds.: *ESA*. Volume 1461 of *Lecture Notes in Computer Science*., Springer (1998) 381-392
- [15] Garey, M., Johnson, D., Stockmeyer, L.: Some simplified np-complete graph problems. *Theoretical Computer Science* 1 (1976) 237-267
- [16] Irving, R.W., Manlove, D.F.: Stable marriage with ties and bounded length preference lists. In: In Proc. ACID, Volume 7 of *Texts in Algorithmics*, 95106, (College Publications)
- [17] Askalidis, G., Immorlica, N., Kwanashie, A., Manlove, D., Pountourakis, E.: Socially stable matchings in the hospitals/residents problem. In Dehne, F., Solis-Oba, R., Sack, J.R., eds.: *Algorithms and Data Structures*. Volume 8037 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg (2013) 85-96

- [18] Gabow, H.N., Tarjan, R.E.: Faster scaling algorithms for network problems. SIAM J. COMPUT 18 (1989) 1013-1036

## AUTHORS

Ashish Shrivastava is a M.S.(By Research) student in Computer Science and Engineering department in Indian Institute of Technology, Madras Chennai. His area of research is Theoretical Computer Science.



C.Pandu Rangan is a Professor in Computer Science and Engineering department in Indian Institute of Technology, Madras Chennai. His area of research is Theoretical Computer Science, Cryptography.



# FAMILY OF 2-SIMPLEX COGNITIVE TOOLS AND THEIR APPLICATIONS FOR DECISION-MAKING AND ITS JUSTIFICATIONS

Yankovskaya Anna<sup>1</sup> and Yamshanov Artem<sup>2</sup>

<sup>1</sup>Tomsk State University of Architecture and Building, Tomsk, Russia  
Tomsk State University of Control Systems and Radioelectronics,  
Tomsk, Russia

National Research Tomsk State University, Tomsk, Russia  
Siberian State Medical University, Tomsk, Russia

ayyankov@gmail.com

<sup>2</sup>Tomsk State University of Control Systems and Radioelectronics,  
Tomsk, Russia

yav@keva.tusur.ru

## **ABSTRACT**

*Urgency of application and development of cognitive graphic tools for usage in intelligent systems of data analysis, decision making and its justifications is given. Cognitive graphic tool "2-simplex prism" and examples of its usage are presented. Specificity of program realization of cognitive graphics tools invariant to problem areas is described. Most significant results are given and discussed. Future investigations are connected with usage of new approach to rendering, cross-platform realization, cognitive features improving and expanding of n-simplex family.*

## **KEYWORDS**

*Cognitive graphics, 2-simplex, 2-simplex prism, decision-making, decision justification, education, e-learning systems, psychosomatic disorders, cognitive modelling*

## **1. INTRODUCTION**

One of the most important and difficult problem in creating the intelligent system for data and knowledge analyzing, and decision-making is the development of a software library for representation of the results in a user friendly and clear view for a wide group of users. It is particularly important when technologies are rapidly changing from data-limited area to data-driven analysis-limited area [1]. The ability to generate big databases of experiments is far ahead from the ability to analyze and visualize these databases. One solution of these problems is usage of the cognitive graphic tools.

An important contribution to the development of the cognitive science was made by R. Axelrod [2], R.G. Basaker [3], D.A. Pospelov [4-6], A.A. Zenkin [7-8], V.F. Khoroshevskiy [9], B.A. Kobrinskiy [10], A.E. Yankovskaya [11-13]. Namely they, who saw the potential of using cognitive tools in the various problem areas, and despite the meager technical capabilities which were available at that time, first steps in the development of the cognitive graphics tools were made, and thereby a new research direction was formed [7, 10]. These tools are very effective for the interpretation of analyzed data and knowledge, decision-making and its justifications for users who are specialized in different problem areas but are not specialists in algorithms of data analysis and knowledge inference which are used in intelligent systems. The cognitive graphic tools are important link between a big amount of data and understanding these data. The application of the cognitive graphic tools makes possible to understand a lot of processes occurring at the lowest level which was not understandable before and revealing of different new regularities and connections between different factors and events in a wide variety of problems and cross-disciplinary areas. The cognitive graphic tools are used in different intelligent systems for information data and knowledge structures analyzing, for revealing regularities of different kinds and decision-making and its justification. They also can be used in the intelligent learning-testing systems for teaching and learning activities optimization, for visualization and forecasting of learning process results and etc. But the development of these tools for each problem area is very time-consuming and expensive. Thus, the cognitive tools which are invariant to different problem areas were developed [14-15]. The specificity of these tools application is the software realization does not aim at a concrete problem area and is realized as visualizing plugins for intelligent instrumental software (IIS) IMSLOG [16] which is used for constructing specific applied intelligent systems. These visualization plugins include the cognitive graphics tools for visualizing information structures, different kinds of regularities, decision-making and its justification etc.

Usage of these tools is not only actual for the parameters analysis of the different non-changing states of objects for decisions justification. It is also actual for analysis and visualization of the dynamic processes. Despite the fact that visualization is not necessary for decision-making, it simplifies the analysis of information and provides possibility for the best decision-making. For example, an intelligent system user in the area of a concrete discipline teaching (a lecturer) considers different aspects of respondents teaching during a time of a real exam or an intelligent system user in the area of the diseases diagnosis (a doctor) considers the previous stages of patients examinations.

The given article continues the research of the cognitive tools application based on the n-simplex [15]. Undoubtful advantages of the cognitive tools based on n-simplex are invariability to the problem areas. It should be pointed out that using the modern technologies for creation of cognitive graphic tools allows to get different cognitive tools: such as desktop application (application for desktop PC), applications for smartphones and pads, WEB application.

Mathematic basis of an object under study representation in n-simplex is described and basis of representation of a process under study in 2-simplex prism is given. Examples of 2-simplex prism application in developed and developing intelligent systems are presented. Further study directions are proposed.

## 2. FAMILY OF 2-SIMPLEX COGNITIVE TOOLS

### 2.1. Cognitive Tool 2-Simplex

Firstly, transformation of features space in patterns space based on the logical-combinatorial methods and properties of n-simplex are suggested in the publication [17]. System of visualization TRIANG for decision-making and its justifications with cognitive graphics [18] is constructed on the base of 2-simplex with usage of the following theorem [17].

*Theorem.* Suppose  $a_1, a_2, \dots, a_{n+1}$  is a set of simultaneously non-zero numbers where  $n$  is the dimension of a regular simplex. Then, there is one and only one point that following condition  $h_1 : h_2 : \dots : h_{n+1} = a_1 : a_2 : \dots : a_{n+1}$  is correct, where  $h_i (i \in 1, 2, \dots, n+1)$  is the distance from this point to  $i$ -th side [17-18].

Coefficient  $h_i (i \in 1, 2, \dots, n+1)$  represents the degree of conditional proximity of the object under study to  $i$ -th pattern [17]. The advantage of this fact is the n-simplex possesses the property of the constancy of the sum of distances ( $h$ ) from any point to each side and the property of ratios preservation  $h_1 : h_2 : \dots : h_{n+1} = a_1 : a_2 : \dots : a_{n+1}$ . Distances  $h_i$  are calculated on the basis of coefficients  $a_i (i \in 1, 2, \dots, n+1)$  and normalization operations from following relations

$$\left\{ \begin{array}{l} H = \sum_{i=1}^n h_i \\ H = A \sum_{i=1}^n a_i \end{array} \right. , \\ \frac{h_1}{a_1} = \frac{h_2}{a_2} = \dots = \frac{h_n}{a_n}$$

where  $A$  – scaling coefficient) by the formula

$$h_i = A \cdot a_i, i \in \{1, 2, \dots, n\}$$

This theorem was used in more than 30 applied intelligent systems and in three instrumental tools of revealing different kind of regularities and making of diagnostic, classifications organization and control decisions and their justification.

The main function of n-simplex is a representation of a disposition of object under study among other objects of a learning sample. Additionally, n-simplex has other useful functions for a decision-making person. One of these functions is a representation of some numerical values, for example, an admissible error of recognition preassigned by the user. Example of 2-simplex is shown on Figure 1.

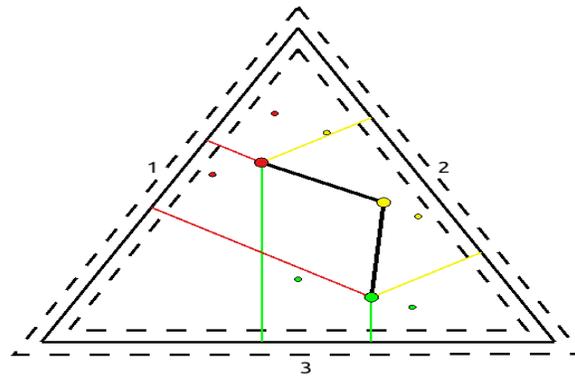


Figure 1. Example of 2-simplex

The sides of 2-simplex (triangle edges) are associated with patterns (classes), circles with big radius are the objects under study and circles with small radius are learning sample objects. The distance from the object to a side is directly proportional to the proximity of the object to the pattern corresponding to the side. The distance for object under study is displayed as color perpendicular lines to 2-simplex sides (red, yellow, green). The color of the object under study or objects from learning sample is mapped to a pattern which is revealed for a specific object. Line segments between objects represent the dynamics of process under study, for example, they can represent changing of a student knowledge level. An object color is mapped with an associated pattern (the nearest pattern or pattern determined by an expert). Digits are mapped with a pattern and are placed on associated sides. It is not a usual working mode because it makes the image more complex, and it is not necessary because the association between the side and the pattern can be determined by a color of the perpendicular line from the point to the side. So, for usual working mode it is suggested to hide these digits. But for the first look or demonstration it can be quite useful.

## 2.2. Cognitive Tool 2-Simplex Prism

The cognitive tool “2-simplex prism” (Figure 2) is based on 2-simplex and represents the triangular regular prism which contains in basics and cuttings 2-simplexes which are corresponded fixed time moments.

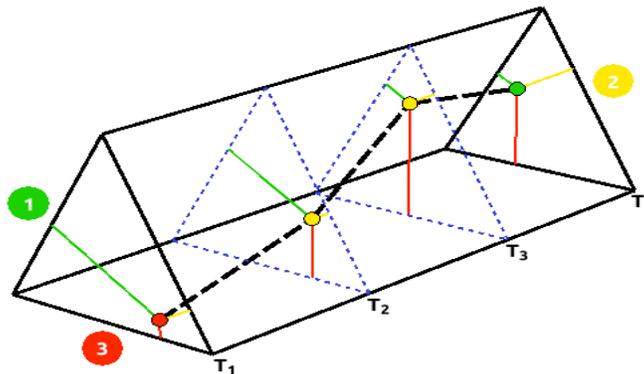


Figure 2. Example of 2-simplex prism

Distance from the base of the prism to  $i$ -th 2-simplex  $h_i'$  corresponds to the fixation moment of object under study features and it is calculated based on the following formula:

$$h_i' = H' \cdot \frac{T_i - T_{\min}}{T_{\max} - T_{\min}},$$

where  $H'$  – length of 2-simplex prism preassigned by a user and corresponded to the study duration,

$T_i$  – timestamp of features fixation of object under study for  $i$ -th examination,

$T_{\min}$  – timestamp of features fixation of object under study for the 1-st examination,

$T_{\max}$  – timestamp of features fixation of object under study for the last examination.

Because 2-simplex prism is based on the 2-simplex description of all 2-simplex objects is also correct for 2-simplex prism.

2-simplex prism allows representing visually as well the dynamic processes as the modeling one or another process, which is necessary for a big amount of problems and cross-disciplinary areas: medicine, economy, genetics, building, radioelectronics, sociology, education, psychology, geology, design, ecology, geo-ecology, eco-bio-medicine etc.

### 3. SOFTWARE IMPLEMENTATION

Software prototypes of the described cognitive tools are implemented with using C# language. Source information for visualization (named listing of  $n$ -simplex or LNS) is a JavaScript code which describes  $n$ -simplex: objects under study, objects of learning sample objects, links between points and other parameters that are necessary for tuning a size of  $n$ -simplex, point of view, type of transformation etc. All described cognitive graphic tools implemented as two libraries: first for 2D visualization (2-simplex and 3-simplex unfoldings) and second for 3D visualization (3-simplex, 2-simplex prism). They have different features, capabilities and visualization parameters but can use one universal LNS. Both  $n$ -simplex library parse and visualize any LNS which describe only objects and which do not describe visualization parameters. If it is necessary to set additional visualization parameters for any library and keep it compatible with another library must be used special language constructions. The libraries visualize  $n$ -simplex only from LNS located in memory. Earlier there was a function to visualize  $n$ -simplex from LNS located in file, but it was excluded because such functionality is not supported on all platforms (for example, java-script for web client) and this functionality not necessary for most intelligent systems and quite trivial in realization. Library “Jint” was used for a LNS parsing. The second version of LNS language was implemented, but the new language is also simple for a code generation as before. Additionally new version of LNS language has more powerful and clearer syntax for creating and editing  $n$ -simplexes by user. The output of the library is a bitmap image visualized by GDI+ library. The fragment of the universal LNS language for a simple 3-simplex is given below.

```
try { setView(0); } catch (ex) { }
try { setTransform(2); } catch (ex) { }
try { setViewPort(15, 80); } catch (ex) { }

var size = 200,
```

```

delta = 40,
colors = ["#E01B1B", "#E0841B", "#F7F307", "#07F70B"],
color = "#000",
dashPattern = [1];

addTetraedron(color, 3, dashPattern, size);

addIJK(color, 2, dashPattern, size, [1, 2, 4, 8], colors);
addIJK(color, 2, dashPattern, size, [8, 4, 2, 1], colors);

addPoint(colors[0], 6, "Circle", size, [1, 2, 4, 8]);
addPoint(colors[1], 6, "Circle", size, [8, 4, 2, 1]);

addPath(color, 4, dashPattern, size, [
  [1, 2, 4, 8],
  [4, 4, 4, 4],
  [8, 4, 2, 1]
]);

```

The visualization and computation algorithms used in both libraries are identical and they are described in the paper [19]. There were various attempts to improve the visualization libraries during this year. The most significant results are the following:

1. An attempt to change a raster renderer to a vector renderer proposed in [18] has failed, because a) partitioning of the objects into smaller entities and sorting them were too time-consuming tasks for rendering in real time to implement the interactive capabilities; b) a developing a vector renderer needs a lot of time and it is too difficult for a quick prototyping.
2. Descriptive language for n-simplex was changed from our custom language to javascript. It was the first step to make library cross-platform and allows integrating our cognitive graphic tools in web applications.
3. Big part of both libraries was universalized and moved to another library. The library contains the code for rendering graphics primitives, algorithms for color transformation, parsing modules etc. This step makes support and development of the libraries simpler and will simplify the porting of the libraries to java-script language which is necessary for an embedding into web-applications.
4. Results of experiments with shaders, OpenGL ES and WebGL were obtained. They show that implementation of all desired functions of planned vector render is also possible with a raster renderer with usage of shaders. Moreover, a shader program is performed using hardware acceleration and all planned tasks can be implemented without losing rich interactive functionality. The implementation of all planned ideas will provide: translucent faces, intelligent layout of signatures, identification of objects at the specified point etc.

The most important advantages of a raster render are rich of graphic abilities (color, layout, complexity etc.) and rich of interactive functions. The most significant disadvantage of a raster render is the impossibility of scaling, but it is not important because n-simplex can be rendered directly in a desired size.

## 4. EXAMPLES OF APPLICATIONS IN DIFFERENT PROBLEM AREAS

A matrix approach to representation of data and knowledge which contains description matrix of objects in space of characteristic features and distinction matrix represented partitioning of objects on equivalence classes for every mechanism of classification are used in intelligent systems (IS) developed by us [14, 20]. The set of all non-regular matrix rows distinction matrix corresponds to set of revealed patterns.

A pattern is a subset of description matrix rows with equal of characteristic features values. In IS based on a logical-combinatory (l-c), a logical-probabilistic (l-p) and a logical-combinatory-probabilistic (l-c-p) methods of test pattern recognition and decision-making with usage of cognitive tools [11, 14, 20] mathematical basis for calculation of conditional proximity coefficients of the object under study to  $i$ -th pattern is proposed. For l-c methods these coefficients are corresponded to a ratio between proximity coefficient object under study  $x$  to a pattern  $k$  and proximity coefficient object into a pattern  $k$ , for l-p and l-c-p these coefficients corresponded with probability to make decision for studied patterns Software implementation of these models for IS include development of corresponded mathematical apparatus for transformation of features space to patterns space which is described in section 2.

### 4.1. Application in Educational Area

This chapter describes visualization of testing knowledge result in learning-testing system with estimation coefficients usage [21]. In learning-testing system developed by us respondent after studying selected discipline, should pass mixed diagnostic test. During solution of this test respondent actions map (RAM) is forming. After completed all the questions, he respondent RAM projected in the set of predetermined valuation coefficients that determine how well the respondent cope with different tasks based on the following abilities (skills):

1. storage and reproduction of the material in unmodified form;
2. the reproduction of the material in modified form;
3. extraction of new knowledge based on the studied material;
4. problem solving, etc.

For example, for development of client-server software system with multimedia capabilities set of evaluated factors reasonable transformations to the following parameters:

1. the solution of problems requiring high concentration;
2. decision-making of non-trivial tasks;
3. fast learning and knowledge of a large number of technologies.

It should be noticed that the different set of these parameters  $(a_1, a_2, a_3)$  can be transformed in same distances  $h_1, h_2, h_3$  in case when sums of  $a_i$  for different sets are equal. So for that and

similar cases it is necessary to introduce the new parameter: a color saturation of point corresponded to the sum of  $a_i$ :  $a_1 + a_2 + a_3$ .

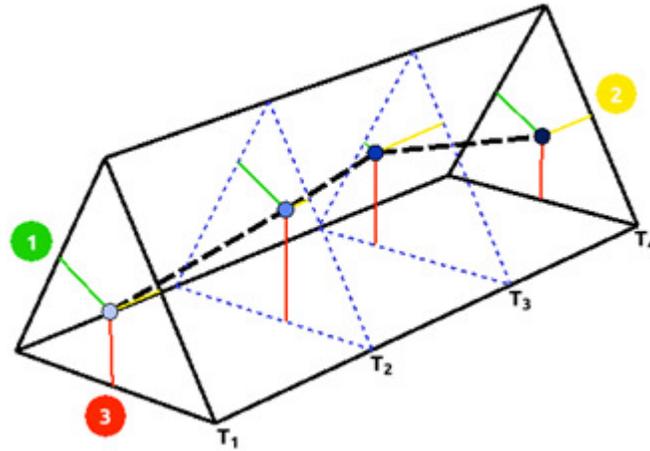


Figure 3. Example of 2-simplex prism usage for learning-testing systems

2-simplex prism allows represent dynamics for ability development a respondent or a group of respondents. But it should be noticed that representation of big group of respondents with usage 2-simplex prism can be too complex and inconvenient.

#### 4.2. Decision Support for Diagnostics and Correction of Psychosomatic Disorders

Recently a number of studies were performed in the field of cognitive behavioral therapy and occupational stress [22–25] in order to identify the persons' psychological state, detect the essential patterns of the problem and correct persons' behavior, providing individual trajectory.

Versatile cognitive tools for identifying the attractiveness of organization are proposed in [26]. Such phenomenon as burnout is thoroughly discussed in [27] and the means of its prevention are considered. The platform for cognitive stimulation, maintenance and rehabilitation for health professionals is proposed in [28]. It allows to define and customize therapeutic intervention programs for cognitive rehabilitation or maintenance, relevant to multiple types of cognitive disorders especially for aged people. In [29] the fact that all processes in psychology happen in time and evolve over time is outlined. Therefore the versatile tool for processes estimation in dynamics is needed for wide range of applications, especially in psychology and medicine. This is especially relevant for person behavior prediction and timely correction in order to prevent their dissatisfaction and turnover intentions [30]. The urge of cognitive testing tools for aged people with disorders of cognition is discussed in [31].

In this paper we propose the cognitive tool designed for the intelligent system for diagnostic and intervention of an organization stress (DIOS) [32]. The following examples of 2-simplex prism usage in area of diagnostic and intervention of organization stress (OS) is given. The system under study is based on the idea of three-stage diagnostic. Our original questionnaire for intervention choosing for every stage (1 - alarm, 2 - resistance, 3 - exhaustion) [33] and fuzzy and threshold logics [34]. The idea of three-stage diagnostic allows in a short period of time to make a differential medical care for a patient with diagnosed OS.

DIOS use special questionnaire for the express-diagnostic of organization stress including a question for three stage OS: stage 1 is an alarm stage, the stage 2 is the resistance stage, stage 3 is

a stage of exhaustion. This questionnaire is based on a Selye conception [35]. 7 features (symptoms) are used for revealing OS in stage 1 and stage 2, 8 features are used in stage 3. Allowed values for each feature are nothing - 0.0, seldom - 0.25, sometimes - 0.5, often - 0.75, constantly - 1.0. Analysis of features values allow to perform an express-diagnostic for revealing OS on every stage.

After the performing of a test system handle patient questionnaire and output diagnostic result of OS for every stage. These results are transferred to a module of visualization and justification. If study of OS recovering dynamics is not required then for justification of result cognitive tools 2-simplex is used. If the dynamics is required then for such purpose 2-simplex prism is used. The represented dynamics of OS diagnostic with a usage of 2-simplex prism allow examine accuracy of revealed diagnosis and selected medical intervention.

2-simplex prism can represent only relation for three patterns, but for some cases more patterns will be necessary. Example of this case is developed by us IS system DIOS which operate with 4 patterns: 3 stages of OS (alarm, resistance, exhaustion) and its absence. Current experiments show that most reasonable way to represent dynamics relating to such cases is usage of additional 2-simplex prism. For DIOS two 2-simplex prisms are used: the first represent object under study relate stages 3-1, the second relate stages 2-1 and OS absence. It should be noticed that both 2-simplex prism are not necessary for all cases. If dynamics of a patient recovering is limited by stages 3-1 or by stages 2-0 and absence it is reasonable to use only one 2-simplex prism.

The follow example of patient dynamics which cannot be limited by one 2-simplex prism is presented. 5 tests for revealing OS were performed. The first 2-simplex prism represents results for 1-4 tests, the second represents for 3-5 tests.

The first 2-simplex prism (Figure 4) represents transformation process from the stage of exhaustion (stage 3) to alarm stage (stage 1).

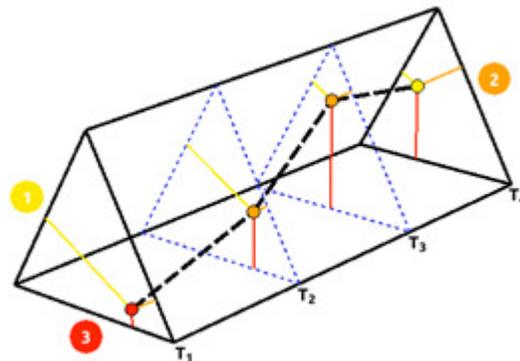


Figure 4. Visualization of tests 1-4 in 2-simplex prism

The first test ( $T_1$ ) reveals a level between the stage of exhaustion and the resistance stage and prepotency of the stage of exhaustion over the resistance stage. The second test ( $T_2$ ) reveals that illness is decreasing from the exhausted stage (pattern 3) to the resistance stage (pattern 2). The third test ( $T_3$ ) reveals that illness is decreased to a level between the resistance stage (pattern 2) and alarm stage (pattern 1). The fourth test ( $T_4$ ) reveals prepotency of the alarm stage (pattern 1).

The second 2-simplex prism (Figure 5) represents the transformation process from the resistance stage (pattern 2) to the absence of stress (pattern 0).

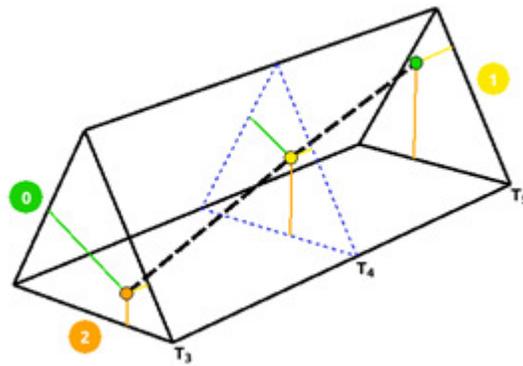


Figure 5. Visualization of tests 2-5 in 2-simplex prism

The fifth test ( $T_5$ ) reveals the absence of the stress organization.

It should be noticed that the cognitive property of color are used in 2-simplex prism to represent dangerous of diagnoses and patterns.

### 4.3. Cognitive Modelling

The cognitive modelling of a decision-making in the artificial intelligent systems is the one of most important directions for creating intelligent systems (IS) in some priority areas of science researches and developing as medicine, psychology, sociology, environmental protection, energetics, systems of transport and telecommunication, control systems etc. Manipulation of some parameters of an object under study and using cognitive tools of decision-making and its justifications in IS we can perform a cognitive modelling base on the different kinds of knowledge representations.

For the cognitive modelling of decision-making only some part of the cognitive tools can be used: these tools should allow to visualize spatial relationship object under study and save sum of distances from any object to patterns (decisions made) and relations between them in process of transformation from space to feature space. Made normalization coefficient of the degree of conditional proximity (for l-c recognition) and set probability of decision making for studied objects (for l-p, l-c-p recognition), representation of object on geometrical figure allows to visualize patterns, see object position among other object from knowledge base and its proximity to a particular pattern. Nowadays in our IS for a cognitive modelling purpose 2-simplex is used, but using of a 2-simplex allows to make a modelling process more clear. The cognitive modeling of decision-making is performed by applying of some actions associated with, for example, using of therapeutic interventions in medicine, the adoption of measures to ensure environmental safety etc. An object answers on some of these actions (some values of some parameters of the object are changing) and their location on 2-simplex is changed. The result of a modeling gives us the dynamic image of an object location and relations between objects in a moment of actions and is represented with the cognitive tools by points which are sequentially connected by a polyline.

The example of prediction and therapeutic intervention with using of 2-simplex prism is presented in Figure 6. After the first patient examination ( $T_1$ ) the diagnosis is revealed - stage 3 (exhausted) of organization stress and strategy of recovering is changed. With usage of mathematical model of a patient and recovery process it is possible to predict progress of patient recovery, which is shown in Figure 6 as polyline of a light-blue color.

After the second examination ( $T_2$ ), a progress of a psychology stress recovering is diagnosed - organization stress is moved from stage 3 to stage 2 (resistance), but a real progress is worse than the predicted progress. At this moment the doctor has two different strategies for continuation of recovering: purple and blue. The doctor uses this cognitive representation of different recovering strategies. He can choose one which gives the best result in the future. At this moment a strategy associated with a polyline of a purple color is more reasonable, because this strategy is applied to a patient. The model predict full patient recovery until the moment of the next examination ( $T_3$ ).

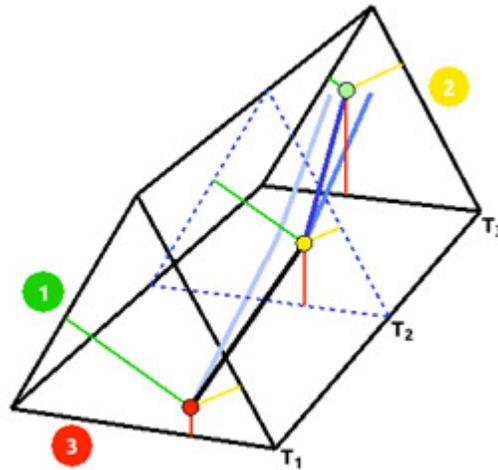


Figure 6. Visualizing of a modeling result in 2-simplex prism.

The cognitive modelling of a decision-making is based on mathematical and computer methods with applying tools form n-simplex family allows to optimize a choice of influence of object under study in accordance with a dynamical model of parameters changing.

Applications of 2-simplex is not limited by the above-mentioned examples.

## 5. CONCLUSIONS

The description of 2-simplex prism and examples of its applications are presented. The most important advantage for the information visualization in 2-simplex prism is the opportunity to analyze in dynamics the object under study over time. It allows the users to make decisions, to justify them and to analyse changes of parameters of object under study.

Application of the cognitive tools can be performed for any problem and cross-disciplinary areas in which it is necessary to make decisions about relation of object under study to one or another pattern (class) in fixed time moment or time interval and justify these decisions. Unlike any of the previously developed cognitive tools based on n-simplex [15, 36], 2-simplex prism allows to study objects dynamically on the time range interested for a user.

Development of cognitive graphics tools invariant to problem areas, their cross-platform realization and their integration in intelligent systems are presented. The implementation of some previously planned steps greatly reduced the time and labor costs for cognitive tools development and improved the human-to-machine interface.

In future we suppose to completely rewrite raster renderer using shaders technology for developing cross-platform realization which can be integrated in web-, desktop- and mobile-applications; to develop interactive features and cognitive properties of described cognitive graphic tools and to expand family of n-simplex.

## ACKNOWLEDGEMENTS

Supported by Russian Foundation for Basic Research, project 13-07-00373a, 13-07-98037-r\_sibir\_a, 14-07-00673 and partially by Russian Humanitarian Scientific Foundation project 13-06-00709.

## REFERENCES

- [1] S. R. Niezgodna, A.K. Kanjarla, S.R. Kalidindi Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data. *Integrating Materials and Manufacturing Innovation* 2013, 2:3 doi:10.1186/2193-9772-2-3.
- [2] R. Axelrod *The Structure of Decision: Cognitive Maps of Political Elites*. Princeton University Press, 1976.
- [3] R.G.Basaker, T.L.Saati *Finite Graphs and Networks: An Introduction with Applications*. Research Analysis Corp., Mc Graw Hill Company, NY-London-Toronto, 1965.
- [4] D.A. Pospelov *Cognitive Graphics is a window into the new world*. *Software products and systems*, 1992, 4-6 (in Russian).
- [5] D.A. Pospelov Ten "hot spots" in research on artificial intelligence. *Intelligent systems (MSU)*, vol. 1(1-4), pp. 47–56, 1996 (in Russian).
- [6] D.A.Pospelov, L.V.Litvintseva How to combine left and right?. *News of Artificial Intelligence*, N2, 1996. (in Russian).
- [7] A.A. Zenkin *Cognitive Computer Graphics*. M.: Nauka, 1991 (in Russian).
- [8] A.A. Zenkin *Knowledge-Generating Technologies of Cognitive Reality*. *News of Artificial Intelligence*, N2, pp. 72-78, 1996. (in Russian).
- [9] V.A.Albu, V.F.Khoroshevskiy *COGR – Cognitive Graphics System, Design, Development, Application*. *Russian Academy of Science Bulletin. Technical Cybernetics - 1990. - № 5* (in Russian).
- [10] B.A. Kobrinskiy Why should we take in account imaginary thinking and intuition in medical expert systems. *Artificial Intelligence – 96. Proceedings of the 5th National Conference with International Participation. Volume II. – Kazan, 1996* (in Russian).
- [11] A.E. Yankovskaya *Decision-making and decision-justification using cognitive graphics methods base on the experts of different qualification*. *Russian Academy of Science Bulletin, Theory and Control Systems*, № 5, pp. 125-126, 1997 (in Russian).
- [12] A. Yankovskaya, D. Galkin *Cognitive Computer Based on n-m Multiterminal Networks for Pattern Recognition in Applied Intelligent Systems*. *Proceedings of Conference GraphiCon'2009. – Moscow.: Maks Press, 2009. – pp. 299-300.*
- [13] A. E. Yankovskaya, D. V. Galkin, G. E. Chernogoryuk *Computer Visualization and Cognitive Graphics Tools for Applied Intelligent Systems*. *Proceedings of the IASTED International Conferences on Automation, Control and Information Technology*, v.1. – 2010. – pp. 249-253.
- [14] A.E. Yankovskaya *Logical tests and means of cognitive graphics*. Publishing house: LAP LAMBERT Academic Publishing, 2011 – 87 c. (in Russian).
- [15] A.E. Yankovskaya, N.M. Krivdyuk *Cognitive graphics tool based on 3-simplex for decision-making and substantiation of decisions in intelligent system*. *Proceedings of the IASTED International Conference Technology for Education and Learning (TEL 2013) – P. 463-469.*
- [16] A.E. Yankovskaya, A.I. Gedike, R.V. Ametov, A.M. Bleikher *IMSLOG-2002 software tool for supporting information technologies of test pattern recognition*. *Pattern Recognition and Image Analysis*, 2003. Vol. 13. No. 4. – P. 650-657.

- [17] A.E. Yankovskaya Transformation of features space in patterns space on the base of the logical-combinatorial methods and properties of some geometric figures. Proceedings of the International Conference Pattern Recognition and Image Analysis: New Information, Abstracts of the I All-Union Conference, Part II, pp. 178-181, Minsk, 1991. (in Russian).
- [18] S.V. Kondratenko, A.E. Yankovskaya System of visualization TRIANG for decision-making justification with cognitive graphics usage. Proceedings of the Third Conference on Artificial Intelligence. Vol. I. - Tver, 1992. - p. 152-155 (in Russian).
- [19] A.V. Yamshanov, N.M. Krivdyuk Specify of software implementation of cognitive graphic tools in intelligent and education systems. Proceedings of Fundamental Science Development XI, Russia, Tomsk, 22–25 April 2014, (ISBN 978-5-4387-0415-7) (in Russian).
- [20] A.E. Yankovskaya An Automaton Model, Fuzzy Logic, and Means of Cognitive Graphics in the Solution of Forecast Problems. Pattern Recognition and Image Analysis. – 1998. – Vol. 8, No. 2. – pp. 154-156.
- [21] A.E. Yankovskaya, Y.A. Shurigin, A.V. Yamshanov, N.M. Krivdyuk Determination of the student knowledge level on the base of a learning course which is presented by a semantic network. Proceedings of Open Semantic Technologies for Intelligent Systems (OSTIS-2015). – Minsk : BSUIR, p. 331-339, 2015 (in Russian).
- [22] Andreea Mateescu, Mihaela Chraif. The relationship between job satisfaction, occupational stress and coping mechanism in educational and technical organizations. Procedia - Social and Behavioral Sciences 187 (2015) 728 – 732.
- [23] Cătălina Dumitrescu. Reducing the self-perceived stress level, heart rate and blood pressure by cognitive behavioral intervention plan in a multinational organization from Romania. Procedia - Social and Behavioral Sciences 187 (2015) 704 – 707.
- [24] Cristiana Catalina Ciceia. Occupational stress and organizational commitment in Romanian public organizations. Procedia - Social and Behavioral Sciences 33 (2012) 1077 – 1081.
- [25] Alan Patchinga, Rick Best. An investigation into psychological stress detection and management in organisations operating in project and construction management. Procedia - Social and Behavioral Sciences 119 ( 2014 ) 682 – 691.
- [26] Che Noriah Othman, Roz Azinur Che Lamin, Nursyuhadah Othman. Occupational Stress Index of Malaysian University. Procedia - Social and Behavioral Sciences 153 (2014) 700 – 710.
- [27] Ruxandra Foloútină, Loredana Adriana Tudorache. Stress management tools for preventing burnout phenomenon at teachers from special education. International Conference on Education and Educational Psychology (ICEEPSY 2012). Procedia - Social and Behavioral Sciences 69 ( 2012 ) 933 – 941.
- [28] José Carlos Teixeira et al. Cognitive stimulation, maintenance and rehabilitation. HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies. Procedia Technology 9 ( 2013 ) 1335 – 1343.
- [29] José Navarro, Robert A. Roe, María I. Artilés. Taking time seriously: Changing practices and perspectives in Work/Organizational Psychology. Journal of Work and Organizational Psychology 31 (2015) 135–145.
- [30] Aharon Tziner, Edna Rabenu, Ruth Radomski, Alexander Belkin. Work stress and turnover intentions among hospital physicians: The mediating role of burnout and work satisfaction. Journal of Work and Organizational Psychology 31 (2015) 207–213.
- [31] Nicole R. Fowler et al. Cognitive testing in older primary care patients: A cluster-randomized trial. Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 1 (2015) 349-357.
- [32] A.E. Yankovskaya, S.V. Kitler, A.V. Silaeva Intelligent system of diagnostics and intervention of organizational stress : its development and testing. Open education 2012. – No 2 (91). – P. 61-69 (in Russian).
- [33] N.A. Kornetov, A.E. Yankovskaya, S.V. Kitler, A.V. Silaeva, L.V. Shagalova about development dynamic of representations about organizational stress and approaches to its evaluation. Fundamental Research №10, 2011 (in Russian).
- [34] Zadeh Lotfi A. Fuzzy Logic, Neural Networks, and Soft Computing. Communications of the ACM. – Vol. 37. – N. 3. – P. 77-84..

- [35] Selye H. A Syndrome Produced by Diverse Nocuous Agents. Nature. – 1936. Vol. 138, – P. 32.
- [36] A.E. Yankovskaya, A.V. Yamshanov, N.M. Krivdyuk Cognitive graphic tools in intelligent learning-testing systems. Proceedings of Open Semantic Technologies for Intelligent Systems (OSTIS-2014). – Minsk : BSUIR, p. 303 -308, 2014 (in Russian).

## AUTHORS

**Professor Anna Yankovskaya** obtained her DSc in Computer Science from the Tomsk State University in Russia. She is currently a head of the Intelligent Systems Laboratory and a professor of the Applied Mathematics Department at Tomsk State University of Architecture and Building, a professor of the Computer Science Department at Tomsk State University, a professor of Tomsk State University of Control Systems and Radioelectronics and a professor Siberian State Medical University. She is the author of more than 600 publications and 7 monographies. Her scientific interests include mathematical foundations for test pattern recognition and theory of digital devices; artificial intelligence, intelligent systems, learning and testing systems, blended education and learning; logical tests, mixed diagnostic tests, cognitive graphics; advanced technology in education.



**Artem Yamshanov** graduated from the Tomsk State University of Control Systems and Radio Electronics in 2012. He is a postgraduate student at the Tomsk State University of Control Systems and Radio Electronics. Research interests: intelligent learning and testing systems, blended education, machine learning, artificial intelligence, intelligent systems and technologies data mining and pattern recognition, cognitive tools and advanced technology in education.



# SEQUENTIAL CLUSTERING-BASED EVENT DETECTION FOR NON- INTRUSIVE LOAD MONITORING

Karim Said Barsim and Bin Yang

Institute of Signal Processing and System Theory,  
University of Stuttgart, Germany

karim.barsim@iss.uni-stuttgart.de

bin.yang@iss.uni-stuttgart.de

## ABSTRACT

*The problem of change-point detection has been well studied and adopted in many signal processing applications. In such applications, the informative segments of the signal are the stationary ones before and after the change-point. However, for some novel signal processing and machine learning applications such as Non-Intrusive Load Monitoring (NILM), the information contained in the non-stationary transient intervals is of equal or even more importance to the recognition process. In this paper, we introduce a novel clustering-based sequential detection of abrupt changes in an aggregate electricity consumption profile with accurate decomposition of the input signal into stationary and non-stationary segments. We also introduce various event models in the context of clustering analysis. The proposed algorithm is applied to building-level energy profiles with promising results for the residential BLUED power dataset.*

## KEYWORDS

*Event detection, change-interval detection, density-based clustering, DBSCAN, non-intrusive load monitoring, NILM, BLUED, energy disaggregation*

## 1. INTRODUCTION

Non-Intrusive Load Monitoring (NILM), also known as electricity disaggregation, is an energy monitoring technique that aims at inferring the energy consumption profiles of individual electrical loads merely from a single or a limited number of aggregate measurement points in a building [1]. Recently, NILM has witnessed a rapidly increasing progress in both academic and commercial research due to its promising applications in energy conservation, activity monitoring [2], dynamic pricing [3], demand forecasting [4], and home automation [5]. Currently, the majority of NILM systems are event-based approaches in the sense that they rely on the detection of abrupt changes occurring in the aggregate signal which indicate state-changes of the monitored appliances. It was observed that events attain distinctive features according to the physical properties of their appliances such as energy storage elements, counter-electromotive force in induction motors, or striking voltages in fluorescent lamps. Features extracted from steady and transient intervals (such as power surges, overshoot currents, decay rate, etc) are utilized in event clustering or classification stages of the disaggregation system. Consequently, a robust detection

and accurate segmentation of such change-intervals is of particular importance for event-based NILM systems.

Basseville and Nikiforov [6] described various detection algorithms from which two approaches have been utilized in event-based NILM systems, namely the Generalized Likelihood Ratio (GLR) test [7, 8] and the CUMulative SUM (CUSUM) filtering [9]. Jin et al. [10] proposed a more robust change-point detection approach based on a Goodness-of-Fit (GoF) test. In addition, various machine learning tools such as kernel clustering [11], Hidden Markov Models (HMM) [12], and Support Vector Machines (SVMs) [13], have been proposed as solutions to address the change point detection problem.

Even though many previous works on NILM proposed utilizing features extracted from the transient intervals, only few event detection approaches consider accurate segmentation of the transient periods for the extraction of more stable transient features [9, 14]. Moreover, many approaches need a probabilistic model for the sample distribution in the stationary segments which is often difficult to obtain from aggregate consumption profile of several, simultaneously operating appliances. The result is that the current event detection algorithms are not robust and fail sometimes provide reliable event-based feature for appliance recognition in practice. In this paper, we propose a novel clustering-based event detection algorithm for event-based NILM systems. In contrast to other event detection algorithms, the proposed approach features accurate segmentation of the input signal into stationary (steady) and non-stationary (transient) segments. Such accurate segmentation is crucial for the extraction of more stable and repeatable features from both transient and steady-state intervals. Moreover, the utilized density-based clustering scheme does not impose any probabilistic models on the sample distribution in either of the stationary segments and supports arbitrarily shaped, weakly stationary segments leading to an enhanced robustness to noise. In addition, the proposed algorithm features a sequential (instead of batch) clustering that is more efficient for real-time NILM systems.

The presented approach is modular in the sense that it can combine any clustering-based event detection algorithm with any event model. For this purpose, we also introduce different event models at different complexity- and robustness-levels. This paper is organized as follows. In section 2, we introduce different event models in the context of spatial and time-series clustering. In section 3, we describe the proposed sequential event detection algorithm in which the Density-Based Spatial Clustering for Applications with Noise [15] is assumed and utilized sequentially in spatial and temporal analysis of the input power signals. Section 4 shows results of application of the proposed algorithm on the publicly available, residential BLUED [16] dataset. Finally, section 5 concludes this paper.

## 2. EVENT MODELS

Event models will be introduced in the order of their increasing coverage of real events, robustness, and complexity.

Let the matrix

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N], \quad \mathbf{x}_n \in \mathbb{R}^l \quad (1)$$

contain a time series of  $N$  consecutive  $l$ -dimensional data samples (feature vectors). Typically,  $\mathbf{x}_n$  contains the measured real  $P$  and reactive  $Q$  powers at time instance  $n$ . Assume that all  $N$  samples have been clustered into  $m$  non-empty, disjoint clusters (sets)  $C_1, C_2, \dots, C_m$ . In addition,

we assume that a noise-aware clustering algorithm assigns un-clustered samples (i.e. outliers or noisy samples) to the set  $C_0$ . Clearly,  $\sum_{i=0}^m |C_i| = N$  where  $|C_i|$  is the cardinality of the cluster  $|C_i|$ . Let

$$y_n = \omega(\mathbf{x}_n) \in \{0, 1, 2, \dots, m\} \quad (2)$$

be the corresponding cluster index of  $\mathbf{x}_n$  (i.e.  $\mathbf{x}_n \in C_{y_n}$ ). We then introduce the following definitions for two metrics of a cluster and three different event models:

**Definition 1:** The *temporal length*  $\text{Len}(C_i)$  of cluster  $C_i$  is defined as the minimum window size that contains all its elements. If

$$\exists u: \mathbf{x}_u \in C_i \text{ and } \mathbf{x}_n \notin C_i \quad \forall (n < u) \quad (3)$$

$$\exists v: \mathbf{x}_v \in C_i \text{ and } \mathbf{x}_n \notin C_i \quad \forall (n > v) \quad (4)$$

Then  $\text{Len}(C_i)$  is defined as

$$\text{Len}(C_i) = v - u + 1 \geq |C_i| \quad (5)$$

Here  $u$  and  $v$  denote the time instances of the first and last samples belonging to  $C_i$ , respectively.

**Definition 2:** The *temporal locality* ratio  $\text{Loc}(C_i)$  of cluster  $C_i$  is defined as

$$\text{Loc}(C_i) = \frac{|C_i|}{\text{Len}(C_i)} \in ]0, 1] \quad (6)$$

The temporal locality ratio is a measure of how a cluster is spreading over time domain. A value of one ( $\text{Loc}(C_i) = 1$ ) refers to the maximum temporal locality where the cluster is represented by a single segment of consecutive observations. This measure is utilized later in the event models as a means to control the amount of noisy samples permitted in the stationary segments.

**Event model  $\mathcal{M}_1$ :** In this event model, a sequence of samples  $\mathbf{X}$  is defined as an event if

- (a) it does not contain any noisy samples (i.e.  $C_0 = \phi$ ),
- (b) it contains two clusters  $C_1$  and  $C_2$  (i.e.  $m = 2$ ),
- (c) both clusters do not interleave (overlap) in the time domain<sup>1</sup>,  
(i.e.  $\exists u : \mathbf{x}_n \in C_1 \quad \forall (n \leq u)$  and  $\mathbf{x}_n \in C_2 \quad \forall (n > u)$ ).

This is the simplest event model without any outliers. It consists of two stationary segments  $\mathbf{X}_{s1} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_u]$  and  $\mathbf{X}_{s2} = [\mathbf{x}_{u+1}, \mathbf{x}_{u+2}, \dots, \mathbf{x}_N]$ . The segment  $\mathbf{X}_t = [\mathbf{x}_u, \mathbf{x}_{u+1}]$  (including the last sample of  $\mathbf{X}_{s1}$  and the first one of  $\mathbf{X}_{s2}$ ) is called the change-interval of the event and  $u$  is the change point. In other words, an event  $\mathcal{M}_1$  is a change interval of length two surrounded by two noise-free weakly stationary segments. This model is valid for switch-off events of most loads as well as switch-on events of resistive ones in a noise-free power signals.

---

<sup>1</sup> For simplicity, and without loss of generality, we assume that the first and second stationary segments of an event are assigned to the cluster sets  $C_1$  and  $C_2$ , respectively.

Figure 1(a) shows an example of a signal segment matching the first event model  $\mathcal{M}_1$  where the scalar samples  $x_n \in \mathbb{R}$  and their corresponding cluster indices  $y_n = \omega(x_n) \in \{1, 2\}$  are plotted over time. The signal represents a step-like event that consists of two stationary segments (red, solid) and a change interval (blue, dashed).

**Event model  $\mathcal{M}_2$ :** A sequence of samples  $\mathbf{X}$  is defined as an event if

- it contains two clusters  $C_1$  and  $C_2$  (i.e.  $m = 2$ ) and the outliers set  $C_0$  is not necessarily empty allowing noisy samples,
- both clusters  $C_1$  and  $C_2$  show a high temporal locality ratio, i.e.  $Loc(C_i) \geq 1 - \epsilon$ , for  $i = 1, 2$
- both clusters do not interleave in the time domain, i.e.  $\exists u, v > u: \mathbf{x}_n \in C_0 \cup C_1 \forall (n < u)$  and  $\mathbf{x}_u, \mathbf{x}_v \in C_1$ , and  $\mathbf{x}_v \in C_0 \cup C_2 \forall (n > v)$  and  $\mathbf{x}_v \in C_2$

Compared with  $\mathcal{M}_1$ , this event model permits noisy samples (i.e. outliers) as well as a lengthy transient interval. This, however, requires the utilization of a noise-aware clustering algorithm. By definition,  $\mathbf{x}_n \in C_0, \forall (u < n < v)$ . In this case, the event contains two stationary segments  $\mathbf{X}_{s1}$  and  $\mathbf{X}_{s2}$  consisting of samples belonging to  $C_1$  and  $C_2$ , respectively, and a change-interval  $\mathbf{X}_t = [\mathbf{x}_u, \mathbf{x}_{u+1}, \dots, \mathbf{x}_{v-1}, \mathbf{x}_v]$ .

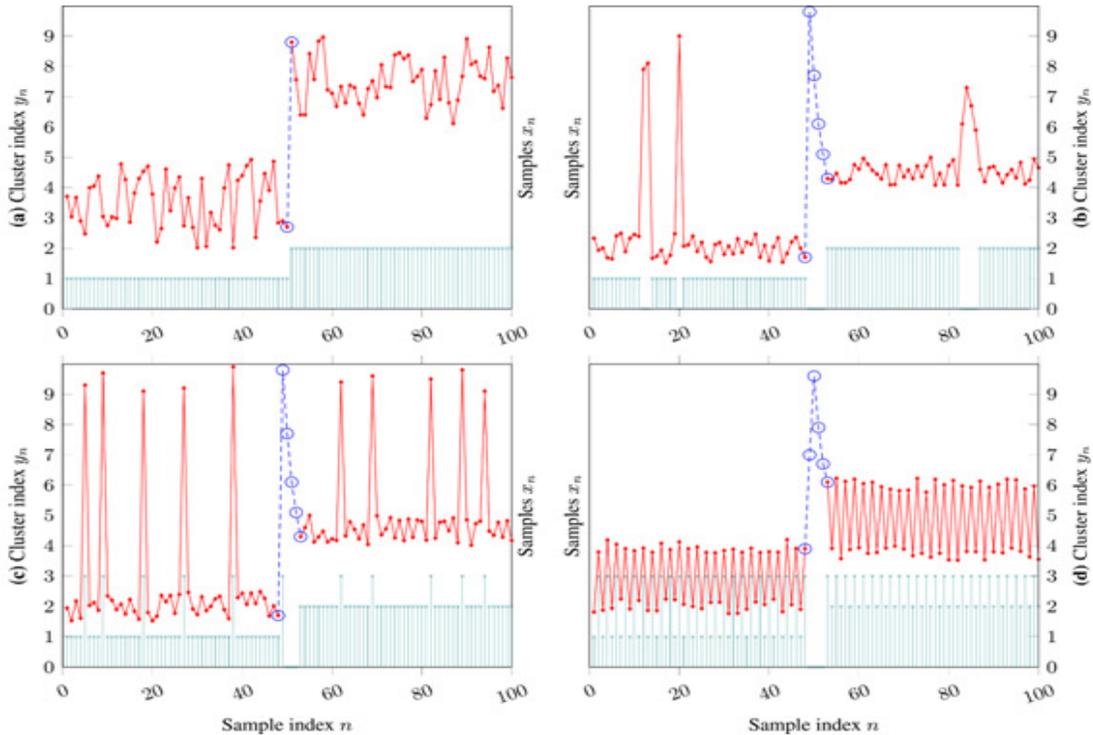


Figure 1: 1-dimensional signals highlighting differences between the three event models. (a) shows a step-like event that is free of both outliers and a transient interval. In (b) random outliers as well as a transient interval are permitted. (c) shows a repeated pattern of spikes that eventually cluster in  $C_3$ . Finally, (d) shows high fluctuations in stationary segments leading to the third cluster  $C_3$  as well. The third event model  $\mathcal{M}_3$  fits all segments, the second event model  $\mathcal{M}_2$  fits only (a) and (b), whereas the first model  $\mathcal{M}_1$  fits only (a).

Figure 1(b) shows an example of a signal segment matching the second event model  $\mathcal{M}_2$  (but not the first one  $\mathcal{M}_1$ ) where the event contains a slower transient interval in a noisy signal. Even though  $\mathcal{M}_2$  is valid for most of the switch-on/off and state-change events within noisy signals, it actually has one implicit assumption on the noise. The assumption that  $m = 2$  (maximally two clusters representing two stationary segments) implies that the noise is random and does not contain a repeated pattern that eventually builds up a cluster when projected to the PQ-plane. This is not always the case as shown in the third example in Figure 1(c).

In the aggregate power signal, some appliances trigger a repeated, sometimes periodic, pattern of high fluctuations or spikes. Such repeated patterns tackle the detection of other actual events. This masking behaviour is resolved in the third event model.

**Event Model  $\mathcal{M}_3$ :** A sequence of samples  $\mathbf{X}$  is defined as an event if

- (a) it contains *at least* two clusters  $C_1$  and  $C_2$  (i.e.  $m \geq 2$ ) and the outliers set  $C_0$  is not necessarily empty,
- (b) clusters  $C_1$  and  $C_2$  show a high temporal locality ratio, i.e.
 
$$Loc(C_i) \geq 1 - \epsilon, \text{ for } i = 1, 2$$
- (c) clusters  $C_1$  and  $C_2$  do not interleave in the time domain, i.e.
 
$$\exists u, v > u: \mathbf{x}_n \notin C_1 \forall (n > u) \text{ and } \mathbf{x}_u \in C_1, \text{ and}$$

$$\mathbf{x}_v \notin C_2 \forall (n < v) \text{ and } \mathbf{x}_v \in C_2$$

In this model, the limitation on the clustering cardinality is released and therefore a repeated noise pattern that eventually results in a wide (temporally wide) cluster would not mask events occurring in the same interval. Similar to  $\mathcal{M}_2$ , the sequence in this model contains two stationary segments  $\mathbf{X}_{s1}$  and  $\mathbf{X}_{s2}$  consisting of samples belonging to  $C_1$  and  $C_2$  respectively, and a change interval consisting of  $\mathbf{X}_t = [\mathbf{x}_u, \mathbf{x}_{u+1}, \dots, \mathbf{x}_{v-1}, \mathbf{x}_v]$ .

Figure 1 (b) and (c) show two event segments fit only by  $\mathcal{M}_3$ . Figure 1(a) shows the simplest event which is fit by all defined models. In Figure 1(b), the transient period as well as the noisy spikes can only be fit by  $\mathcal{M}_2$  and  $\mathcal{M}_3$ . Finally, the repeated noise pattern in Figure 1(c) or high fluctuations in Figure 1(d) only match the last event model  $\mathcal{M}_3$ .

## 2. DETECTION ALGORITHM

The main task of the event detection algorithm is to search for signal segments that match a given event model  $\mathcal{M}_i$ . This is achieved by applying a clustering algorithm on different segments and checking how much each segment matches the model. In all of the three models introduced in section 2, the clustering cardinality  $m$  is not known in advance. Therefore, a utilized clustering algorithm should either be nonparametric or a model order estimation step has to take place beforehand.

In our approach we utilized the commonly used Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [15]. The DBSCAN algorithm (or density-based clustering in general) has several advantages that make it the best candidate for a non-parametric sequential event detection. First, DBSCAN assumes no prior knowledge of the number of clusters. Second, DBSCAN supports arbitrarily shaped clusters with no constraints on their samples' distribution. In addition, DBSCAN is a noise-aware clustering algorithm and, therefore, can be utilized with any of the previously defined event models.

Ideally, the detection algorithm searches the input signal sequentially for segments that match a given event model. However, we control the matching process with a proximity measure that shows how much a segment matches the given model.

**Definition 3:** The model loss between an event model  $\mathcal{M}_i$  and a signal segment  $\mathbf{X}$  is defined as

$$\begin{aligned} \mathcal{L}(\mathcal{M}_i, \mathbf{X}, u, v) = & |\{\mathbf{x}_n: n \leq u \text{ and } \mathbf{x}_n \in C_2\}| + \\ & |\{\mathbf{x}_n: n \geq v \text{ and } \mathbf{x}_n \in C_1\}| + \\ & |\{\mathbf{x}_n: u < n < v \text{ and } \mathbf{x}_n \in C_1 \cup C_2\}| \end{aligned} \quad (7)$$

where  $u$  and  $v$  are the indices of the first and last sample of the change-interval, respectively. In the case of  $\mathcal{M}_1$  where  $v = u + 1$ , the last term in Equation 7 becomes zero regardless of  $u$ .

The model loss function counts the number of samples that need to be corrected (i.e. reassigned to a different set  $C_j$  of the clustering structure) in order for the segment  $\mathbf{X}$  to match the event model  $\mathcal{M}_i$ . The lower the loss, the more the signal segment matches the event model.

The proposed detection algorithm can then be presented as to two sub-tasks, the forward detection step which is the main process for finding an event, and the backward reduction step that is responsible for a more accurate segmentation.

In the forward detection step, new samples are received one at a time and inserted into the clustering space. Upon insertion of a new sample, the clustering indices are updated and the model loss is re-estimated. Once a match is encountered (i.e. the model loss is zero or less than a predefined threshold  $\lambda$ ), a detection is declared with the current change point  $u$  of the matched segment and the change-interval  $\mathbf{X}_t = [\mathbf{x}_u, \mathbf{x}_{u+1}, \dots, \mathbf{x}_{v-1}, \mathbf{x}_v]$  where  $\mathbf{x}_v$  is the first sample of the second stationary segment.

Once an event is declared, the backward reduction step begins. In this step, samples are removed from the clustering space in a First-In-First-Out (FIFO) fashion while updating the clustering structure upon each deletion and re-estimating the model loss. The reduction ends by the last sample that satisfies the matching condition (i.e. if that sample is deleted, the segment will no longer matches the event model within the predefined threshold loss  $\lambda$ ). The complete detection algorithm can be described as follows. Given an event model  $\mathcal{M}_i$

1. Receive new sample  $\mathbf{x}_{N+1}$  and append it to  $\mathbf{X}$
2. Update the clustering vector  $\mathbf{y}$  and the clustering structure  $\{C_j\}_{j=1}^m$
3. Check  $\mathcal{L}(\mathcal{M}_i, \mathbf{X}, u, v) \leq \lambda$  for all  $u, v$ , if not satisfied, go to step (1)
4. Declare event detection with change-interval  $\mathbf{X}_t = [\mathbf{x}_u, \mathbf{x}_{u+1}, \dots, \mathbf{x}_{v-1}, \mathbf{x}_v]$  and change-point is  $u$  where  $u$  and  $v$  result in the minimum model loss between  $\mathcal{M}_i$  and the current segment  $\mathbf{X}$  (i.e.  $\text{argmin}_{u,v} \mathcal{L}(\mathcal{M}_i, \mathbf{X}, u, v)$ ).
5. Delete oldest sample  $\mathbf{x}_1$  from the segment
6. Update the clustering vector  $\mathbf{y}$  and the clustering structure  $\{C_j\}_{j=1}^m$
7. Check  $\mathcal{L}(\mathcal{M}_i, \mathbf{X}, u, v) \leq \lambda$  for all  $u, v$ , if satisfied, go to step (5)
8. Re-insert last sample and declare current segment  $\mathbf{X}$  as a balanced event.

After each detection, the process restarts from the first sample of the second stationary segment  $\mathbf{x}_v$ . The main objective of the backward reduction step is to extract balanced stationary segments (i.e.  $|C_1| \approx |C_2|$ ) around the transient interval. Balanced segments lead to more stable steady-state features as well as an enhanced robustness to missed detections (i.e. false negatives).

## 2. EXPERIMENTS AND RESULTS

The proposed event detection approach has been evaluated on different power datasets among them is the Building-Level fully labelled Electricity Disaggregation (BLUED) dataset [16]. In the following, we show the results of applying the event detection algorithm with event model  $\mathcal{M}_3$  and the DBSCAN clustering scheme on the BLUED dataset. We only show evaluation of detection results. Evaluation of the accuracy of transient interval segmentation and the stability of extracted features is beyond the scope of this paper.

Table 1 shows the event detection results on the real and reactive power signals from the BLUED dataset. BLUED include aggregate measurements from a two-phase residential building (phase A and B) and each is evaluated separately. True Positives (TP) is the number of successful detections, False Positives (FP) is the number of detections that do not correspond to actual events, while False Negatives (FN) is the number of missed events. Finally, False Positive Percentage (FPP), precision, recall, and the F1-score measures are defined as

$$\text{FPP} = \frac{FP}{E} \quad (8)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{recall} = \frac{TP}{TP + FN} \quad (9)$$

$$F_1 - \text{score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (10)$$

where  $E$  is the number of events. Results show highly precise detection rates where the number of false positives is relatively low in both phases. It is also observed that, noise in the second phase (phase B) still masks a relatively large number of events.

Table 1. Event detection results on BLUED [16] dataset.

	Phase A	Phase B	Total
Number of events $E$	892	1609	2501
Number of detections	874	1176	2050
True Positives (TP)	867	1097	1964
False Positives (FP)	7	79	86
False Negatives (FN)	25	512	537
FPP	0.78%	4.91%	3.44%
precision	<b>99.20%</b>	<b>93.28%</b>	<b>95.81%</b>
recall (TPR)	<b>97.20%</b>	<b>68.18%</b>	<b>78.53%</b>
$F_1$ -score	<b>98.19%</b>	<b>78.78%</b>	<b>86.31%</b>

### 3. CONCLUSIONS

We introduced a novel clustering-based approach for sequential event detection. The proposed algorithm features accurate segmentation of the stationary and non-stationary intervals for more stable feature extraction, support of arbitrarily shaped stationary segments with no prior assumptions on their sample distribution, and more robustness to noise as well as parameter variations.

### REFERENCES

- [1] G. W. Hart, "Nonintrusive appliance load monitoring", in proceedings of the IEEE: vol.80, no.12, pp. 1870-1891, Dec. 1992. doi:10.1109/5.192069
- [2] A. I. Cole and A. Albicki, "Data extraction for effective non-intrusive identification of residential power loads", in proceedings of the Instrumentation and Measurement Technology Conference (IMTC) 1998 IEEE: vol.2, pp.812-815, May 1998. doi:10.1109/IMTC.1998.676838
- [3] S. Drenker and A. Kader, "Nonintrusive monitoring of electric loads", in Computer Applications in Power, IEEE: vol.12, no.4, pp.47-51, Oct 1999. doi:10.1109/67.795138
- [4] M. El Hachemi Benbouzid, "A review of induction motors signature analysis as a medium for faults detection", IEEE Transactions on Industrial Electronics, vo.47, no.5, pp.984-993, Oct 2000.
- [5] S. N. Patel, T. Robertson, J. A. Kientz, M. S. Reynolds, and G. D. Abowd, "At the Flick of a Switch: Detecting and Classifying Unique Electrical Events on the Residential Power Line", in UbiComp 2007: Ubiquitous Computing, vol.4717, pp.271-288, 2007.
- [6] M. Basseville and I. V. Nikiforov. Detection of Abrupt Changes: Theory and Application. Prentice Hall, 1993.
- [7] M. Berges, E. Goldman, L. Soibelman, H. S. Matthews, and K. Anderson, "User-centred non-intrusive electricity load monitoring for residential buildings", Journal of Computing in Civil Engineering, vol.25, no.1, 2011.
- [8] K. D. Anderson, M. E. Berges, A. Oceanu, D. Benitez, and J. M. F. Moura, "Event Detection for Non-Intrusive Load Monitoring", in IECON 2012, 38th Annual Conference on IEEE Industrial Electronics Society, October 2012.
- [9] K. N. Trung, E. Dekneuevel, B. Nicolle, and O. Zammit, "Event Detection and Disaggregation Algorithms for NIALM System", in the 2nd International Non-Intrusive Load Monitoring (NILM) Workshop, Jun 2014.
- [10] Y. Jin, E. Tebekaemi, M. Berges, and L. Soibelman, "A time-frequency approach for event detection in nonintrusive load monitoring" in proceedings of the Signal Processing, Sensor Fusion, and Target Recognition, Orlando, Florida, USA, 2011.
- [11] M. Volpi, D. Tuia, G. Camps-Valls, and M. Kanevski, "Unsupervised Change Detection With Kernels", in Geoscience and Remote Sensing Letters, IEEE: vol.9, no.6, pp.1026-1030, Nov. 2012. doi: 10.1109/LGRS.2012.2189092
- [12] M. Luong, V. Perduca, and G. Nuel, "Hidden Markov Model Applications in Change-Point Analysis", arXiv Journal, preprint arXiv:1212.1778v1, 2012.

- [13] G. L. Grinblat, L. C. Uzal and P. M. Granitto, “Abrupt change detection with one-class time-adaptive support vector machines”, *Expert Systems with Applications Journal*, vol. 40, pp. 7242–7249, 2013.
- [14] S. B. Leeb, S. R. Shaw, and Jr. Kirtley, J. L. “Transient event detection in spectral envelope estimates for nonintrusive load monitoring”, *IEEE Transactions on Power Delivery*, 10(3):1200–1210, July 1995.
- [15] M. Ester, H. P. Kriegel, J. Sander, and X. Xu, “A density based algorithm for discovering clusters in large spatial databases with noise” in proceedings of Knowledge Discovery and Data mining (KDD), 1996.
- [16] K. Anderson, A. Ocneanu, D. Benitez, D. Carlson, A. Rowe, and M. Berges, “BLUED: a fully labeled public dataset for Event-Based Non-Intrusive load monitoring research”, in proceedings of the 2nd Knowledge Discovery and Data mining (KDD) Workshop on Data Mining Applications in Sustainability (SustKDD), Beijing, China, August 2012. URL: <http://nilm.cmubi.org/>

*INTENTIONAL BLANK*

# HOL, GDCT AND LDCT FOR PEDESTRIAN DETECTION

Sanaa Tayb<sup>1</sup>, Youssef Azdoud<sup>1</sup>, Aouatif Amine<sup>1</sup>, Bouchra Nassih<sup>1</sup>, Hanaa Hachimi<sup>1</sup> and Nabil Hmina<sup>1</sup>

<sup>1</sup>LGS, National Schools of Applied Sciences,  
Ibn Tofail University, Kenitra, Morocco

sanaa.tayb@gmail.com, azdoud.youssef@univ-ibntofail.ac.ma,  
amine\_aouatif@univ-ibntofail.ac.ma, bouchra.nassih@gmail.com,  
hanaa.hachimi@univ-ibntofail.ac.ma, hmina@univ-ibntofail.ac.ma

## **ABSTRACT**

*In this paper, we present and analyze different approaches implemented here to resolve pedestrian detection problem. Histograms of Oriented Laplacian (HOL) is a descriptor of characteristic, it aims to highlight objects in digital images, Discrete Cosine Transform DCT with its two version global (GDCT) and local (LDCT), it changes image's pixel into frequencies coefficients and then we use them as a characteristics in the process. We implemented independently these methods and tried to combine it and used there outputs in a classifier, the new generated classifier has proved it efficiency in certain cases. The performance of those methods and their combination is tested on most popular Dataset in pedestrian detection, which are INRIA and Daimler.*

## **KEYWORDS**

*Pedestrian detection, HOL, DCT local, DCT Global, Classification, SVM, Inria Person, Daimler.*

## **1. INTRODUCTION**

In the last years, pedestrian detection and tracking systems have been an important major and a subject of investigation in the vision system studies. Various approaches and methods have been proposed to fulfil the task of detection. It is an application that intend to determine the pedestrian position in an image or a scenery, its importance is noticeable in so many areas of our everyday life, such as traffic safety, video surveillance, car safety, or automatic driver-assistance systems in vehicles.

A robust pedestrian detection system have be able to detect people in different circumstances: different shapes and scales, huge amount of variation in the poses, and some cases of partially occluded pedestrian or whose parts blend with complex background sometimes. Therefore, we proposed a new method and tested it on two different databases with completely different set of contents that considered all the above circumstances.

This paper impart an efficient approach to deal with the pedestrian detection problem, for this purpose we propose a method based on three major tasks: features extraction using HOL, GDCT and LDCT, dimensionality reduction of feature vectors and finally the classification employing the SVM classifier tested out with many kernel functions.

The primary idea behind this work is to prove the effectiveness of the combination of two approaches on the performance of the pedestrian detection rate.

This paper is organized as follows: we will start by displaying the pre-processing techniques in section 3 and 4, where HOL and its important properties is presented, next both Global and Local DCT are introduced. The experimental results, discussion and the main experiments were conducted on INRIA person and Daimler sets. Finally yet importantly, a comparison between different approaches results is given in section 6.

## 2. RELATED WORK

Pedestrian detection remains a two-class classification problem that aims to classify an input image as a positive if a pedestrian occurs in the image or a negative if no pedestrian is found in. Previous approaches have been proposed in literature based on so many features extraction methods, we note VJ [14] as short for the Viola and Jones one of the first descriptor proposed and used to detect objects efficiently in real-time images, it was also trained using the AdaBoost classifier in [11] and merged both motion and appearance information to spot a walking person, Hog descriptor displayed by Dalal and Triggs [4] it is a feature descriptor used in computer vision and image processing in the purpose of detecting objects, Hol [2] and so many others that can be uncouncted in [3]. In [15] David et al. represent a new approach using Haar wavelets and edge orientation histograms fed into an Adaboost classifier to generate a good pedestrian detection compared to the approach results displayed in [4], in which Dalal and Triggs used support vector machine with histogram of oriented gradient. In the context of detecting object in the paper [16] an integral image was adopted to allow the Adaboost classifier to compute quickly the features judged critical. Then their generated classifier is combined in cascade way in order to discard quickly some background regions while concentrating computations on interest object regions. In [17] Gavrilu and Munder presents a multi-cue vision system which use cascade modes that analyse visual criteria to narrow down the search space to perform detection, similarly [18] exhibit a framework that counts four detectors trained individually to get four components of the human body, the results are then combined and classified either a “person” or a “nonperson.”, besides the work in [19] is an evaluation of the performance of descriptors which compute local interest regions and it proposes an extension of SIFT descriptor according to their results. The most popular approach for improving detection quality is to increase/diversify the features computed over the input image, by having richer and higher dimensional representations, the classification task becomes somehow easier, enabling improved results. To that matter a large set of feature types have been explored: edge information [5] [6] [7] [8], color information [6] [12], texture information [13], local shape information co-variance features [9], among others. More and more diverse features have been shown to systematically improve performance.

### 3. HOL

#### 3.1. Generalities

The Histogram of oriented Laplacien proposed in [2] is an image descriptor similar to Hog descriptor. HOG is a set of characteristic descriptors, which serves to detect objects in digital images. It was introduced by Dalal and Triggs [4] and has contributed in so many works in literature in the field of features extraction. The Hog descriptor is based on the principle that the shape and the appearance of an object is described by the gradient intensity distribution and orientation of edges. In fact, the implementation of these descriptors starts by cutting the image in adjacent small areas called cells. For each cell, the histogram of the gradient orientations is commuted for each pixel. Hence, combining all those calculated histograms gives the HOG descriptor. To polish the given results, the histograms are normalized by calculating a measure of the intensity of a set of cells called blocks. This measure is used to normalize all the cells in a block. The process of building a HOG descriptor for a given image is described as follows:

- First; the input image is divided into small spatial regions corresponding to the cells.
- Then, the input image is divided into small special regions corresponding to the cells.
- For each cell, accumulate a local histogram of gradient directions or edge orientation in every pixels in the cell
- The normalization is applied on local histogram to reduce the effect of illumination variance.

This can be done by accumulating a measure of these local histograms over regions called blocks. A block is a set of cells and then use the results to normalize all cells in the block.

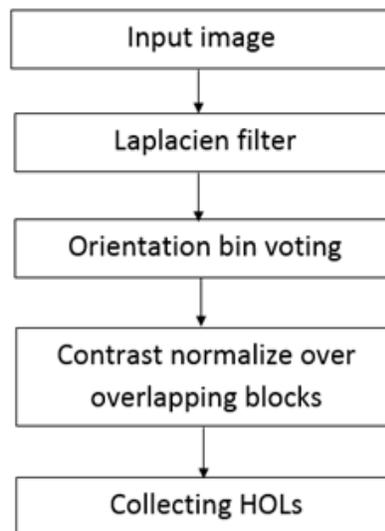


Figure 1. Operating principle of HOL

The HOL descriptor (see Figure 1), is based on the processing of the HOG descriptor where the gradient filter was replaced by the Laplacien one.

## 4. THE DISCRETE COSINE TRANSFORM

GDCT (Global Discrete cosine transform) and LDCT (Local discrete cosine transform) for pedestrian detection were two transforms introduced and used for pedestrian detection. Unlike other transforms, the DCT attempts to represent each image by a set of features called DCT coefficients where most of the visually significant information about the image is concentrated in just a few coefficients of the DCT, most of the time we combine the coefficient extraction to coefficient reductions to reduce the computation time.

### 4.1. Generalities

The Discrete Cosine Transform helps separate the image into parts (or spectral sub-bands) of differing importance (with respect to the image's visual quality). The DCT is similar to the discrete Fourier Transform since it transforms a signal or image from the spatial domain to the frequency representation.

As known in an image, most of the energy is concentrated in the lower frequencies, so if we transform an image into its frequency components and neglect the higher frequency coefficients, we can reduce the amount of data needed to describe the image without sacrificing too much image quality. Ahmed, Natarajan, and Rao (1974) [1] first introduced the DCT in the early seventies.

The general equation for a 2D (N x M image) DCT is defined by the following equation:

$$F(u, v) = \left(\frac{2}{N}\right)^{1/2} \left(\frac{2}{M}\right)^{1/2} \sum_{j=0}^{M-1} \sum_{i=0}^{N-1} \Lambda(i, j) \cos \left[ \frac{\pi u}{2N} (2i + 1) \right] \cos \left[ \frac{\pi v}{2M} (2j + 1) \right] f(i, j) \quad (1)$$

Where

$$\Lambda(i) \begin{cases} 1/\sqrt{2} & \text{for } u = 0 \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

$$\Lambda(j) \begin{cases} 1/\sqrt{2} & \text{for } v = 0 \\ 1 & \text{otherwise} \end{cases} \quad (3)$$

And

f (i,j) is the 2D input sequence

The basic operation of the DCT is as follows:

- 1) The input image is N by M;
- 2) f(i,j) is the intensity of the pixel in row i and column j;
- 3) F(u,v) is the DCT coefficient in row k1 and column k2 of the DCT matrix;
- 4) For most images, much of the signal energy lies at low frequencies; these appear in the upper left corner of the DCT;
- 5) The output array of DCT coefficients contains integers; these can range from -1024 to 1023;
- 6) It is computationally easier to implement and more efficient to regard the DCT as a set of basic functions which given a known input array size (8 x 8) can be precomputed and stored. This involves simply computing values for a convolution mask (8 x8 window)

that get applied (sum  $m$  values  $\times$  pixel the window overlap with image apply window across all rows/columns of image). The values are simply calculated from the DCT formula.

In Natural images, the majority of the DCT energy is concentrated on low frequencies, so each image retain a set of low frequency features and high frequency features (edges) which are rarely encountered.

#### 4.2. Feature reduction

The advantages of DCT are based on the fact that most images have sparse edges. Therefore, most blocks contain primarily low frequencies, and can be represented by a small number of coefficients without significant precision loss. In general, the first DCT coefficients contain most of the information about the global statistics of the processed block, which is necessary to achieve high classification accuracy. Based on that, the idea of reducing the dimension of the DCT coefficient vector length of GDCT and LDCT is eventual.

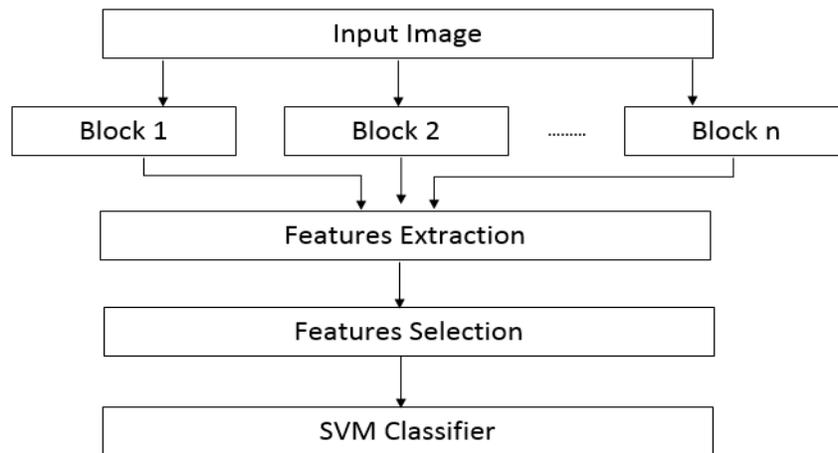


Figure 2. The principle of applying GDCT

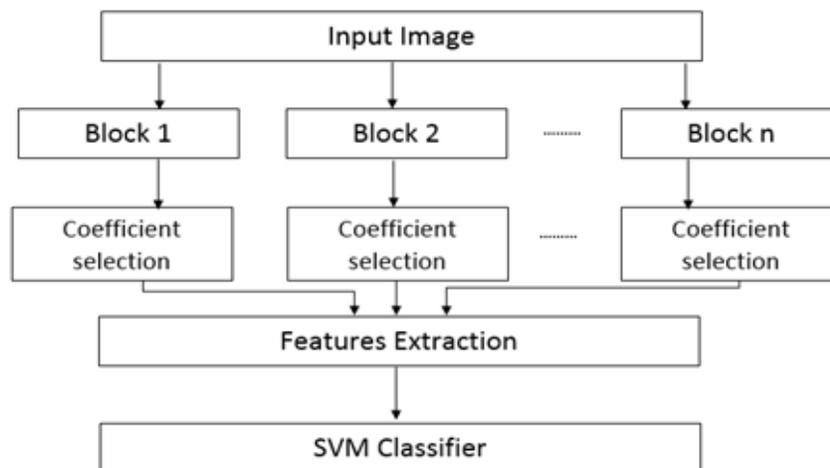


Figure 3. The principle of applying LDCT

## 5. SUPPORT VECTOR MACHINE

The theory of SVMs is from statistics, which is proposed by Vapnik [10]. The basic principle of SVM is finding the optimal linear hyper plane in the feature space that maximally separates the two target classes. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different classes memberships. The SVM is a state-of-the-art classification method and it is widely used in supervised classification in machine learning applications. Apart from its simplicity and accuracy, the SVM classifier is popular because of the ease with which it can deal with high-dimensional data. It performs binary classification by defining a hyper-plane that classifies the input data into two classes (Pedestrian) and (Non pedestrian). SVM uses kernel functions to classify data in a high dimension space, where the separation is performed much easily. The separation is given by the equation:

$$F(x) = \sum_{i=1}^n (\alpha_i y_i K(x_i, x)) + b \quad (4)$$

## 6. EXPERIMENTAL RESULTS AND DISCUSSION

Much of the progress in the past few years has been driven by the availability of challenging public datasets. In this paper, the proposed approach has been mainly applied on INRIA person along with the Daimler datasets, and simulated on the platform of Matlab.

### 6.1. Databases

Multiple pedestrian datasets have been collected over the years and used for testing different approaches in the field of pedestrian detection or tracking.

The INRIA person dataset is very popular in the Pedestrian Detection community, both for training detecting and reporting results. It helped drive recent advances in pedestrian detection and remains one of the most widely used databases despite its limitations. This dataset was collected as part of research work on detection of upright people in images and video.

In addition to the Inria person dataset, our approach has been tested on the Daimler pedestrian.



Figure 4. Samples of images from the INRIA person Database

The Daimler pedestrian dataset contains a collection of pedestrian and non-pedestrian images. It is made available for download over the internet for benchmark purposes, in order to advance research on pedestrian classification.



Figure 5. A sample of DC database

## 6.2. Proposed Method

In this paper, we propose a new approach for the pedestrian detection; it is based on the concatenations of the HOL characteristic vector respectively with the GDCT and the LDCT coefficients.

In this paper, all the images of the Inria person dataset were re-sized to square 290x290 pixels for the GDCT application, and to 288x288 pixels so they can be an integral number of 8x8 pixels per block for the LDCT application, as a result we end up with 1296 blocks of 8x8 pixels. Based on the principle of the features reduction represented in the Figure 2 and Figure 3, we reduced the number of coefficients for both GDCT and LDCT, the obtained vectors are then concatenated to the HOL feature vector and then fed into a Bi-class SVM to classify the input data as a pedestrian ID or not. LIBSVM (a library for Support Vector Machines) and different kernel functions were tested here to improve detection, by reducing the miss rate, and increasing the accuracy rate Table 1.

Table 1. Test on the Inria person.

	Kernel	Coefficients	Accuracy (%)	Miss Rate (%)
HOL	RBF	-	77.42	18
GDCT	Polynomial	84100 coef / Image	76.91	23
	Polynomial	40000 coef / Image	<b>76.91</b>	<b>23</b>
LDCT	Polynomial	64 coef / Block	76.91	23
	Polynomial	16 coef / Block	<b>76.91</b>	23
GDCT+HOL	Polynomial	-	76.91	23
	Linear	-	<b>78.60</b>	<b>21</b>
LDCT+HOL	Polynomial	-	32.08	86
	Linear	-	<b>32.56</b>	<b>83.03</b>

Many combination of SVM parameter were tested, the best performances were obtained using the Polynomial kernel function, and we reduced the number of the coefficients per images since we obtain better performance. The outcome of the concatenation of the GDCT and the HOL descriptor is the increase of the detection rate by 2 % and the decrease of the miss rate compared to the application of HOL or GDCT as a feature descriptor on the Inria person dataset each one alone .

As for the application GDCT of the LDCT on the Daimler, we had to rescale the images respectively to 39x39 pixels which give us a total of 1296 coefficients per image, and to 32x32 pixels for the LDCT that provide us with 16 blocks of 8x8 pixels per image, and each block enumerate 64 coefficients, the deferent results are summed up on Table 2

Table 2. Test on the DC database

	Kernel	Coefficients	Accuracy (%)	Miss Rate (%)
HOL	RBF	-	53.01	23
GDCT	Polynomial	1296 coef / Image	65.5	18
LDCT	Polynomial	64 coef / Block	67.27	19.7
GDCT+HOL	Polynomial	-	<b>65.48</b>	<b>18</b>
	Linear	-	65.12	39.80
LDCT+HOL	Polynomial	-	32.36	64.7
	Linear	-	<b>48.59</b>	<b>52.37</b>

### 6.3. Comparison

This section provides a comparative study of our approaches and multiple well-known techniques for pedestrian detection applied on Inria person database and the DC pedestrian data set.

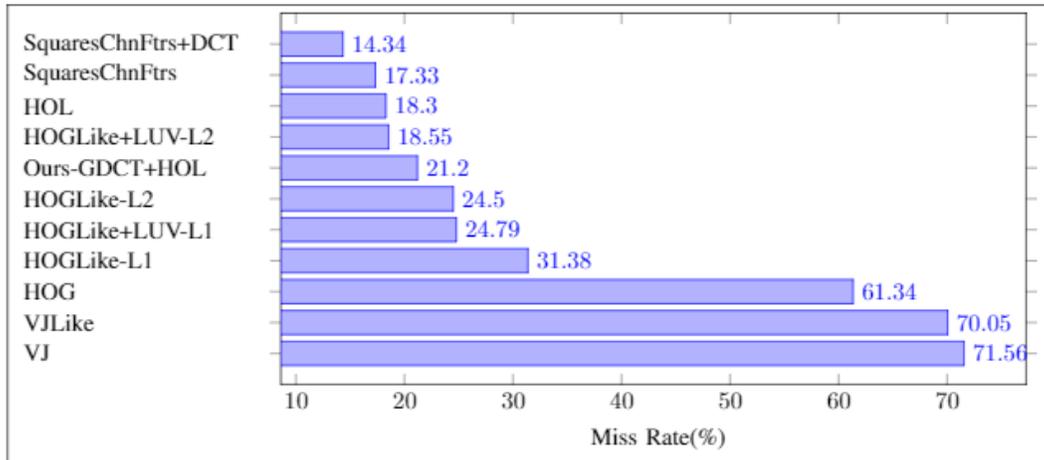


Figure 6. Effect of features on detection performance Inria Dataset

Classification performance of our proposed method is evaluated using the bar plot represented on the Figure 6, which plot the miss rate.

The miss rate is a measure that calculate the proportion of false negative rate miss detection of a pedestrian in an image, it is considered in so many research works as a procedure that shows the effectiveness of different features in the detection applied on the Inria person database.

As we have shown experimentally, our approach brings several advances when compared to commonly used and existing procedures; the reduction of the number of coefficients calculated by the mean of GDCT and LDCT, since they don't capture discriminative information, combining it to the HOL descriptor reduces the Miss rate, and places our approach in a decent position compared to the HOG [4] combined to the SVM classifier and VJ [11], the most used descriptors in the field of pedestrian detection.

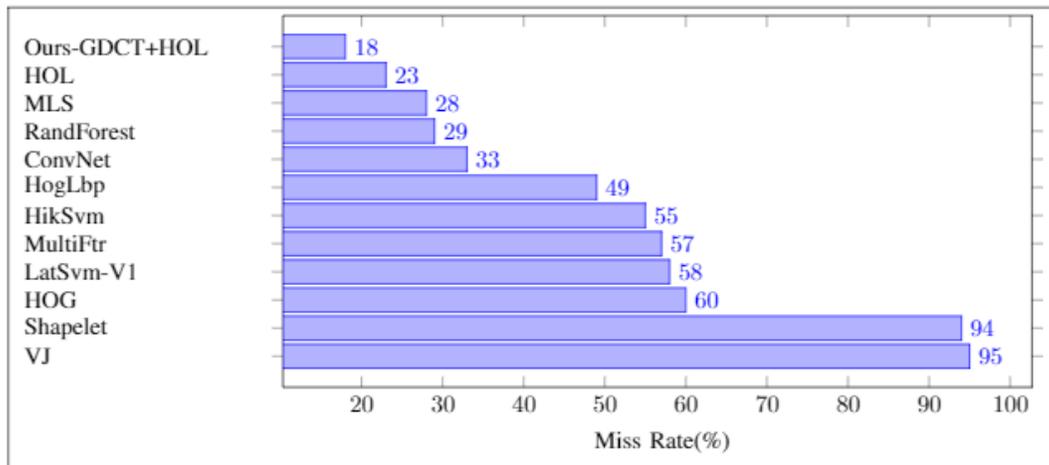


Figure 7. Effect of features on detection performance Daimler Dataset

The Daimler dataset is grayscale unlike the other datasets. As such, only a subset of the detection algorithms were applicable (those that did not rely on color information). The Figure 7 shows that the proposed feature selection proposed in this article outperform different methods proposed out there and applied on the Daimler database.

## 7. CONCLUSION

In this study, we putted forward an extended approach to the proposed method GDCT and LDCT applied for the pedestrian detection by combining these methods to HOL approach.

Before the combination of the two separate approaches, we dug the possibility of features reduction in both cases of GDCT and LDCT, reducing the number of studied coefficients and keeping only the more significant ones, then concatenating them to the HOL coefficient.

We showed experimentally that our proposed method for detecting pedestrian improves over some recent techniques proposed in the article [2] as well as the earlier method of [4], from which HOL seeks out its idea.

This article is an in depth comparison of the proposed method and the state of the art methods for the pedestrian detection, applied and verified on the Inria Person and Daimler databases; regarding the classification methods, we used the SVM classifier and tested various kernel functions and kept the most performing ones.

## REFERENCES

- [1] N. Ahmed, T. Natarajan and K. Rao, (1974). 'Discrete Cosine Transform', IEEE Transactions on Computers, vol. -23, no. 1, pp. 90-93.
- [2] R. Benenson, M. Omran, J. Hosang and B. Schiele, (2015). 'Ten Years of Pedestrian Detection, What Have We Learned?', Computer Vision - ECCV 2014 Workshops, pp. 613-627.
- [3] A. Costea and S. Nedeveschi.,( 2014). 'Word Channel Based Multiscale Pedestrian Detection without Image Resizing and Using Only One Classifier', 2014 IEEE Conference on Computer Vision and Pattern Recognition.

- [4] N. Dalal and B. Triggs,( 2005). 'Histograms of Oriented Gradients for Human Detection',IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05).
- [5] P.Dollár, Z.Tu, P.Perona and S.Belongie, (2009). 'Integral Channel Features', 2009, In BMVC, volume 2,page 5.
- [6] J. Jaewhan Lim, P.Dollar and C.Lawrence Zitnick III, (2013). 'Learned mid-level representation for contour and object detection', US Patent Application pp.13/794,857.
- [7] P. Luo, Y. Tian, X. Wang and X. Tang, (2014). 'Switchable Deep Network for Pedestrian Detection', IEEE Conference on Computer Vision and Pattern Recognition.
- [8] Y.Azdoud ,A. Amine , N.Alioua and M.Rziza, (2015 ). 'Precollision detection system for pedestrian safety based on hol'. IEEE/ACS 12th International Conference on Computer Systems and Applications (AICCSA).
- [9] S. Paisitkriangkrai, C. Shen and A. Hengel, (2013). 'Efficient Pedestrian Detection by Directly Optimizing the Partial Area under the ROC Curve', IEEE International Conference on Computer Vision.
- [10] V. Vapnik, (1998). Statistical learning theory. New York: Wiley.
- [11] P. Viola, M. Jones and D. Snow, (2005). 'Detecting Pedestrians Using Patterns of Motion and Appearance', Int J Comput Vision, vol. 63, no. 2, pp. 153-161.
- [12] S. Walk, N. Majer, K. Schindler and B. Schiele, (2010). 'New features and insights for pedestrian detection', IEEE Computer Society Conference on Computer Vision and Pattern Recognition.
- [13] X. Wang, T. Han and S. Yan, (2009). 'An HOG-LBP human detector with partial occlusion handling', IEEE 12th International Conference on Computer Vision.
- [14] P. Viola and M. Jones, (2004). 'Robust Real-Time Face Detection', International Journal of Computer Vision, vol. 57, no. 2, pp. 137-154.
- [15] D. Gerónimo, A. López, D. Ponsa and A. Sappa, (2007). 'Haar Wavelets and Edge Orientation Histograms for On-Board Pedestrian Detection', Pattern Recognition and Image Analysis, pp. 418-425.
- [16] P. Viola and M. Jones, 'Rapid object detection using a boosted cascade of simple features', (2001). Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR.
- [17] Gavrila and S. Munder, 'Multi-cue Pedestrian Detection and Tracking from a Moving Vehicle', (2006). International Journal of Computer Vision, vol. 73, no. 1, pp. 41-59.
- [18] A. Mohan, C. Papageorgiou and T. Poggio, 'Example-based object detection in images by components', (2001), IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 4, pp. 349-361.
- [19] K. Mikolajczyk and C. Schmid, 'A performance evaluation of local descriptors', (2005). IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 10, pp. 1615-1630.

# AN ADVANCED TOOL FOR MANAGING FUZZY COMPLEX TEMPORAL INFORMATION

Aymen Gammoudi<sup>1,2</sup>, Allel Hadjali<sup>2</sup> and Boutheina Ben Yaghlane<sup>3</sup>

<sup>1</sup>LARODEC/ISGT, 92, Boulevard 9 Avril 1938, 1007 Tunis, Tunisia  
aymen.gammoudi@ensma.fr

<sup>2</sup>LIAS/ENSMA, 1 Avenue Clement Ader, 86960 Futuroscope Cedex, France  
allel.hadjali@ensma.fr

<sup>3</sup>IHECT, University of Carthage,  
IHEC-Carthage Presidence, 2016 Tunis, Tunisia  
boutheina.yaghlane@ihec.rnu.tn

## ABSTRACT

*Many real-life applications need to handle and manage time pieces of information. Allen temporal relations are one of the most used and known formalisms for modelling and handling temporal data. This paper discusses a novel idea to introduce some kind of flexibility in defining such relations between two fuzzy time intervals. The key concept of this approach is a fuzzy tolerance relation conveniently modelled. Tolerant Allen temporal relations are then defined using the dilated and the eroded intervals of the initial fuzzy time intervals. By leveraging some particular fuzzy indices to compare two fuzzy time intervals, this extension of Allen relations is integrated in the Fuzz-TIME system developed in our previous works.*

## KEYWORDS

*Temporal Relations, Fuzzy Time Intervals, Tolerance Relation, Dilation, Erosion, Temporal Databases.*

## 1. INTRODUCTION

Among advanced databases we find in real world applications, temporal databases which manage temporal aspects (time, date ...) of the data they contain are often encountered. In such applications, temporal information is often perceived or expressed in an imprecise/fuzzy manner [1]. For example, periods of global revolutions are characterized by beginnings and endings naturally gradual and ill-defined (such as "well after early 20" or "to late 30"). Unfortunately, and to the best of our knowledge, there is not much work devoted to querying/handling fuzzy/imprecise information in a temporal databases context. On the contrary, in Artificial Intelligence field, several works exist to represent and handle imprecise or uncertain information in temporal reasoning (see for instance, Dubois et al. [2], Schockaert S. and De Cock [3], Badaloni and Giacomini [4]).

As mentioned above, only very few studies have considered the issue of modeling and handling flexible queries over regular/fuzzy temporal databases. In [5] [6], an approach that integrates bipolar classifications to determine the degree of satisfaction of records, is proposed. It relies on using both positive and negative imprecise and possibly temporal preferences. But this approach is still unable to model complex temporal relationships and cannot be applied in historical temporal databases (for instance, the user may request one time period but reject a part of this period, when specifying the valid time constraint in the query). Deng et al. [7] have proposed a temporal extension to an extended ERT model to handle fuzzy numbers. They have specified a fuzzy temporal query which is an extension from the TQuel language and they have introduced the concepts of fuzzy temporal in specification expressions, selection, join and projection. Tudorie et al. [8] have proposed a fuzzy model for vague temporal terms and their implication in queries' evaluation. Unfortunately, this approach does not allow to model a large class of temporal terms (such as: just after and much before). In [9], Galindo and Medina have proposed an extension of temporal fuzzy comparators and have introduced the notion of dates in Relational Databases (RDB) by adding two extra precise attributes on dates (VST, VET). Recently, in [10] we have proposed an extension, named *TSQLf*, of *SQLf* language [11] by adding the time dimension. *TSQLf* language allows for expressing user queries involving fuzzy criteria on time. It is founded on the fuzzy extension of Allen temporal relations already proposed in [2].

Unfortunately, all the above approaches consider (fuzzy) temporal relations only between regular time intervals (i.e., their lower and upper bounds are crisp instants). While in real world applications, time intervals are often described by ill-defined bounds to better capture the vagueness inherent to the available pieces of time information.

This paper is a step towards dealing with that issue. It proposes an extension of Allen temporal relations to compare fuzzy time intervals. This extension relies on a particular tolerance relation that allows associating a fuzzy time interval with two nested fuzzy time intervals (i.e., the dilated and the eroded intervals). Based on these two nested intervals, in order to introduce some softness in comparing fuzzy temporal entities, tolerant Allen relations are defined. Particular fuzzy indices are used for the purpose of tolerant Allen relations computation.

The paper is structured as follows. In Section 2, we provide some background necessary to the reading of the paper and a critical related work. In section 3, tolerant Allen relations modeling and their handling are explicitly discussed. Then in section 4, we describe how we have integrated this extension in our *Fuzz-TIME* system. Section 5 concludes the paper and sketches some lines for future work.

## 2. BACKGROUND AND RELATED WORK

The purpose of this section is manifold. We begin by recalling Allen temporal relation, and then we recall some fuzzy comparators of interest. Finally, we present the dilation and erosion operations on fuzzy sets. This section is mainly browsed from [2].

### 2.1. Allen Temporal Relations

Allen [12] has proposed a set of mutually exclusive primitive relations that can be applied between two temporal intervals. These relationships between events are usually denoted by *before* ( $\prec$ ), *after* ( $\succ$ ), *meets* (m), *met by* (mi), *overlaps* (o), *overlapped by* (oi), *during* (d),

contains (di), starts (s), started by (si), finishes (f), finished by (fi), and ( $\equiv$ ). Their meaning is illustrated in Table 1 (where  $A = [a, a']$  and  $B = [b, b']$  are two time intervals with  $a$  and  $a'$  (respectively  $b$  and  $b'$ ) represent the two bounds of  $A$  (respectively  $B$ ), with  $a < a'$ ).

### 2.2. Fuzzy comparators

In this section, we recall two fuzzy comparators expressed in terms of difference of values. Such comparators capture approximate equalities and graded inequalities.

#### Approximate Equalities and Graded Inequalities

An *approximate equality* between two values, here representing dates, modeled by a fuzzy relation  $E$  with membership function  $\mu_E$  ( $E$  stands for "equal"), can be defined in terms of a distance such as the absolute value of the difference. Namely,

$$\mu_E(x, y) = \mu_L(|x - y|)$$

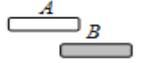
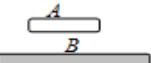
For simplicity, fuzzy sets and fuzzy relations are assumed to be defined on the real line. Approximate equality can be represented by

$$\forall x, y \in \mathbb{R}$$

$$\mu_E(x, y) = \mu_L(|x - y|) = \max\left(0, \min\left(1, \frac{\delta + \varepsilon - |x - y|}{\varepsilon}\right)\right)$$

$$= \begin{cases} 1 & \text{if } |x - y| \leq \delta \\ 0 & \text{if } |x - y| > \delta + \varepsilon \\ \frac{\delta + \varepsilon - |x - y|}{\varepsilon} & \text{otherwise} \end{cases}$$

Table 1. Allen Relations.

Relation	Inverse	Signification	Relations between bounds
$A < B$	$B > A$		$b > a'$
$A m B$	$B mi A$		$a' = b$
$A o B$	$B oi A$		$b > a \wedge a' > b \wedge b' > a'$
$A d B$	$B di A$		$a > b \wedge b' > a'$
$A s B$	$B si A$		$a = b \wedge b' > a'$
$A f B$	$B fi A$		$a > b \wedge b' = a'$
$A \equiv B$	$B \equiv A$		$a = b \wedge a' = b'$

where  $\rho$  and  $\epsilon$  are respectively positive and strictly positive parameters which affect the approximate equality. With the following intended meaning: the possible values of the difference  $a - b$  are restricted by the fuzzy set  $L = (-\delta, \delta, \epsilon, \epsilon)$ <sup>1</sup>. In particular  $a E(0) b$  means  $a = b$ .

Similarly, a *more or less strong inequality* can be modeled by a fuzzy relation  $G$  ( $G$  stands for "greater"), of the form

$$\mu_G(x, y) = \mu_K(x - y)$$

In the following, we take

$$\forall x, y \in \mathbb{R}$$

$$\mu_G(x, y) = \mu_K(x - y) = \max\left(0, \min\left(1, \frac{x - y - \lambda}{\rho}\right)\right) = \begin{cases} 1 & \text{if } x > y + \lambda + \rho \\ 0 & \text{if } x \leq y + \lambda \\ \frac{x - y - \lambda}{\rho} & \text{otherwise} \end{cases}$$

We assume  $\rho > 0$ , i.e.  $G$  more demanding than the idea of "strictly greater" or "clearly greater".  $K = (\lambda, \lambda + \rho, +\infty, +\infty)$  is a fuzzy interval which gathers all the values equal to or greater than a value fuzzily located between  $\lambda$  and  $\rho$ .  $K$  is thus a fuzzy set of positive values with an increasing membership function. See Figure 1. According to the values of parameter  $\lambda + \rho$  the modality, which indicates how much larger than  $b$  is  $a$ , may be linguistically labelled by "Slightly", "moderately", "much". In a given context  $G(0)$  stands for '>'. In a given context  $L(0)$  stands for '='.

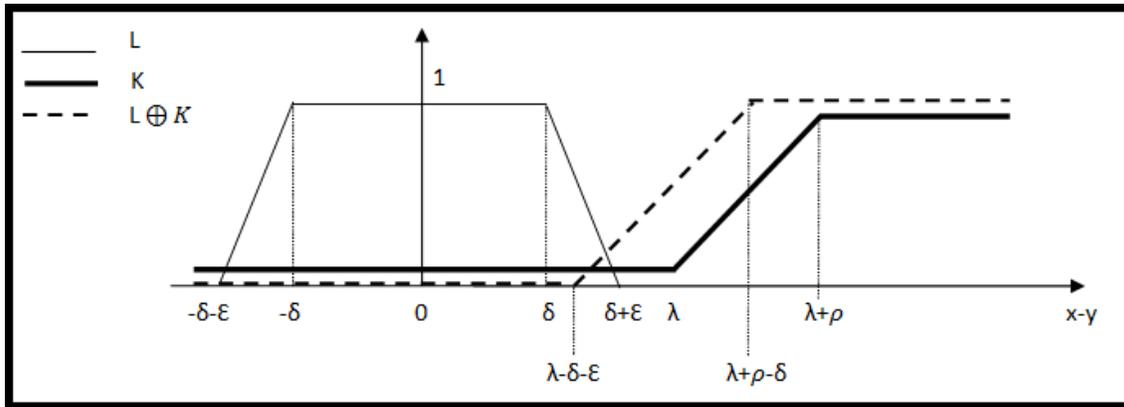


Figure 1. Modeling "approximate equality" and "graded strict inequality".

### 2.3. Dilation and Erosion Operations

Let us consider a fuzzy set  $A$  representing a time interval, and an *approximate equality* relation  $E(L)$ .  $A$  can be associated with a nested pair of fuzzy sets when using the parameterized relation  $E(L)$  as a tolerance relation [2]. Indeed,

<sup>1</sup>  $A = (A, B, a, b)$  stands for the trapezoidal membership function (t.m.f.) of the fuzzy set  $A$  where  $[A, B]$  (resp.  $[A-a, B+b]$ ) is the core (resp. support).

- ❖ one can build a fuzzy set of temporal instants close to  $A$  such that  $A \subseteq A^L$ . This is the dilation operation,
- ❖ one can build a fuzzy set of temporal instants close to  $A$  such that  $A_L \subseteq A$ . This is the erosion operation.

### 2.2.1. Dilation operation

Dilating the fuzzy set of temporal instants  $A$  by  $L$  will provide a fuzzy set  $A^L$  defined by

$$\mu_{A^L}(r) = \sup_s \min(\mu_{E[L]}(s, r), \mu_A(s)) \tag{1}$$

$$= \sup_s \min(\mu_L(r - s), \mu_A(s)) \tag{2}$$

$$= \mu_{A \oplus L}(r) \tag{3}$$

Hence,

$$A^L = A \oplus L$$

where  $\oplus$  is the addition operation extended to fuzzy sets [13].  $A^L$  gathers the elements of  $A$  and the elements outside  $A$  which are somewhat close to an element in  $A$ . See Figure 2.

One can easily check that the fuzzy set of temporal instants  $A^L$  is less restrictive than  $A$ , but still semantically close to  $A$ . Thus,  $A^L$  can be viewed as a relaxed variant of  $A$ . In terms of t.m.f., if  $A = (a, a', \alpha, \alpha')^2$  and  $L = (-\delta, \delta, \epsilon, \epsilon)$  then  $A^L = (a - \delta, a' + \delta, \alpha + \epsilon, \alpha' + \epsilon)$ , see Figure 2.

#### Example:

If  $A = (15, 19, 2, 1)$  and  $L = (-1, 1, 0.5, 0.5)$  then

$$A^L = (14, 20, 2.5, 1.5)$$

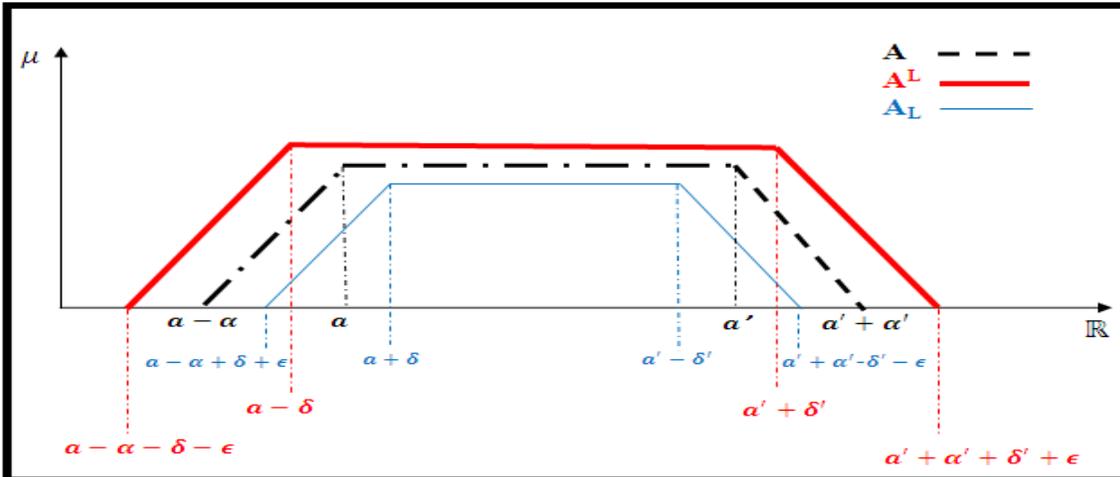


Figure 2. Dilated and eroded time intervals of a fuzzy set of temporal instants  $A$ .

<sup>2</sup> with  $[a, a']$  (resp.  $[a-\alpha, a'+\alpha]$ ) represents the core (resp. support) of  $A$ .

### 2.2.2. Erosion operation

Let  $L \oplus X = A$  be an equation where  $X$  is the unknown variable. Solving this equation has extensively been discussed in [14]. It has been demonstrated that the greatest solution of this equation is given by  $\bar{X} = A \ominus (-L) = A \ominus L$  since  $L = -L$  and where  $\ominus$  is the extended Minkowski subtraction defined by [11]:

$$\mu_{A \ominus L}(r) = \inf_s \left( \mu_L(r - s) \Rightarrow_{\mathbb{T}} \mu_A(s) \right) \quad (4)$$

where  $\mathbb{T}$  a t-norm, and  $\Rightarrow_{\mathbb{T}}$  is the R-implication induced by  $\mathbb{T}$  and defined by  $\Rightarrow_{\mathbb{T}}(u, v) = \sup\{\lambda \in [0, 1], \mathbb{T}(u, \lambda) \leq v\}$  for  $u, v \in [0, 1]$ . We make use of the same t-norm  $\mathbb{T}$  ( $= \min$ ) as in the dilation operation which implies that  $\Rightarrow_{\mathbb{T}}$  is the so-called Gödel implication.

Let  $(E[L])_r = \{s, \mu_{E[L]}(s, r) > 0\}$  be the set of elements that are close to  $r$  in the sense of  $E[L]$ . Then, the above expression can be interpreted as the degree of inclusion of  $(E[Z])_r$  in  $A$ . This means that  $r$  belongs to  $A \ominus L$  all the elements  $s$  that are close to  $r$  are  $A$ . Hence, the inclusion  $A \ominus L \subseteq A$  holds. This operation is very useful in natural language to intensify the meaning of vague terms. Now, eroding the fuzzy set  $A$  by  $L$  results in the fuzzy set  $A_L$  defined by

$$A_L = A \ominus L.$$

The fuzzy set  $A_L$  is more precise than the original fuzzy set  $A$  but it still remains not too far from  $A$  semantically speaking. If  $A = (a, a', \alpha, \alpha')$  and  $L = (-\delta, \delta, \epsilon, \epsilon')$  then  $A_L = A \ominus L = (a + \delta, a' - \delta, \alpha - \epsilon, \alpha' - \epsilon)$  provide that  $\alpha \geq \epsilon$  and  $\alpha' \geq \epsilon$ . Figure 2 illustrates this operation.

#### Example:

If  $A = (15, 19, 2, 1)$  and  $L = (-1, 1, 0.5, 0.5)$  then

$$A_L = (16, 18, 1.5, 0.5)$$

In the crisp case,  $A \ominus L = [a, a'] \ominus [-\delta, \delta'] = [a + \delta, a' - \delta']$  (while  $A \oplus Z = [a - \delta, a' + \delta']$ ). One can easily check that the following proposition holds:

**Proposition 1.** Using the t.m.f. of  $A^L$  and  $A_L$  given above, we have:

- $(A^L)_L = (A_L)^L = A$
- $(A^L)^L = A \oplus 2L$
- $(A_L)_L = A \ominus 2L$

### 2.4. Related Work

Temporal information is often perceived or expressed in a vague and imprecise manner. Here we discuss some works related to the modeling and handling of imperfections in time both in Artificial Intelligence and Databases fields.

The treatment of imprecise or uncertain information in temporal reasoning has been addressed for a longtime. Dubois and Prade in [15] discuss the approximate reasoning on fuzzy dates and fuzzy

intervals in the framework of possibility theory. Guesgen et al. [16] propose fuzzy Allen relations viewed as fuzzy sets of ordinary Allen relationship taking into account a neighborhood structure.

Fuzzy sets, which play a key role in the modeling of flexible constraints, have also been used in different approaches based on constraints for temporal reasoning. Qian and Lu [17] have studied several propagation strategies for the treatment of fuzzy rules temporal networks. Barro et al. [18] have proposed a generalization based on fuzzy sets of time constraint and used possibility measures to verify the consistency's degree of a fuzzy temporal constraint network. Godo and Vila [19] have defined a temporal logic approximated based on the integration of fuzzy constraints in a logical language. The inference system is based on specific rules treating fuzzy constraints and proposals degrees of certainty. Dubois et al. [20] have proposed a possibilistic temporal logic that a formula which is associated with a fuzzy set of dates when the formula is more or less certainly true.

On the other hand, only few works have been proposed for dealing with imperfect data in databases. Billiet et al. [5] have proposed an approach that integrates bipolar classifications to determine the degree of satisfaction of records by using both positive and negative imprecise and possibly temporal preferences. But this approach is still unable to model complex temporal relationships and not applied in historical temporal databases (for instance, the user may request one time period but reject a part of this period, when specifying the valid time constraint in the query). Tudorie et al. [8] have proposed a fuzzy model for vague temporal terms and their implication in queries' evaluation. Unfortunately, this approach does not allow to model a large number of temporal terms (such as: just after and much before). Galindo and Medina [9] have proposed an extension of temporal fuzzy comparators and have introduced the notion of dates in Relational Databases (RDB) by adding two extra precise attributes on dates (VST, VET). The most disadvantage of this approach that cannot support some sophisticated queries that need a step of reasoning before processing. However, some comparators might not be in full agreement with the intuitive semantics underlying the notion of the temporal relations that refer to.

However, in [10] we have proposed an extension, named *TSQLf*, of *SQLf* language [11] by the addition of the dimension time. *TSQLf* language allows for expressing user queries involving fuzzy criteria on time. It is founded on the fuzzy extension of Allen temporal relations already proposed in [2]. Recently, in [21] we have proposed a first step to introduce some flexibility in defining such relations between two fuzzy time intervals. This idea presents an extension of Allen temporal relations based on a particular tolerance relation that allows associating a fuzzy time interval with two nested intervals (i.e., the dilated and the eroded intervals).

Unfortunately, all the above approaches consider (fuzzy) temporal relations only between regular time intervals (i.e., their lower and upper bounds are crisp instants). While in real world applications, time intervals are often described by ill-defined bounds to better capture the vagueness inherent to the available pieces of time information. In this context (i.e., bounds of temporal intervals are ill-defined), not much work exist in the literature. Except, the works done by Nagypal and Motik [22], Ohlbach [23] and Schockaert et al. [3]. Nagypal and Motik [22] have defined a temporal model based on fuzzy sets. This model extends Allen relations with fuzzy time intervals (ITFs). An ITF means a temporal interval with bounds defined in an imprecise way (for example, "the period from the late 20s to the early 30s" is an ITF with the following semantics (1928, 1933, 2, 2)). Nagypal and Motik have introduced a set of auxiliary operators on intervals such as, for example, the operator taking an interval *I* and built intervals containing all

instants which are before initial time  $I$ . Then, fuzzy counterparts of these operators have been defined on the ITFs. Extended Allen relations with ITFs were introduced using the fuzzy operators. Note that the composition of these relations was not discussed by the authors. Schockaert et al. [3] also proposed a generalization of Allen relations with ITFs. This generalization allows handling classical relations between imprecise events (such as, "Roosevelt died before the start of the Cold War"), and also imprecise nature relations (such as, "Roosevelt died just before the start of the Cold War "). The key notion used in this approach is the concept of fuzzy orders on time (as, for example, the fuzzy order that expresses how a moment  $a$  is much smaller than a moment  $b$ ). These orders represented by parameterized fuzzy relations are applied to the gradual bounds ITFs to define, for example, the degree  $bb^{<<}(A, B)$  (expressing how the beginning of an ITF  $A$ , defined in an imprecise manner, is before the end of an ITF  $B$  defined in an imprecise way as well). In [4], Badaloni and Giacomini have developed a fuzzy extension of the algebra of intervals, named  $IA^{fuz}$  where degrees of preference (belonging to  $[0, 1]$ ) are attached to each atomic relation between two classical time intervals.

### 3. TOLERANT ALLEN RELATIONS

The purpose of this section is twofold. We start by modeling tolerant Allen relations based on dilation and erosion operations. Then, we present the comparing indices of two fuzzy intervals to compute to what extent a tolerant Allen relation is satisfied.

#### 3.1. Modeling

Using the dilatation and erosion operations defined above, we can provide the basis for defining a tolerance-based extension of Allen relations. For instance, the tolerance-based temporal relation corresponding to *meet* Allen relation writes:

$A \text{ toler-meets}(L) B$  would correspond to  $A_L \text{ before } B_L$  and  $A^L \text{ overlaps } B^L$

The statement  $A \text{ toler-meets}(L) B$  means that the temporal relation between the two (crisp/fuzzy) temporal intervals  $A$  and  $B$  is perceived as a variant of *meet* relation thanks to some tolerance expressed by the indicator  $L$ . This is a human perception which is often encountered in real world applications (such as in managing historical temporal data, planning and scheduling, natural language understanding, etc.).

The *toler-meet* relation gathers a class of Allen relations (i.e., *before* and *overlaps*) applied on the eroded and dilated time intervals corresponding to the original time intervals  $A$  and  $B$ . This boils down to compute the traditional Allen relations, *before* and *overlaps*, on fuzzy temporal intervals  $A^L$  and  $A_L$  (since  $A^L$  and  $A_L$  are fuzzy sets as shown in the previous subsection).

Let  $A = (a, a', \alpha, \alpha')$  be a fuzzy time interval with  $\tilde{a} = (a, a, \alpha, 0)$  and  $\tilde{a}' = (a', a', 0, \alpha')$  the two fuzzy bounds of validity of  $A$ . One can write  $A = (\tilde{a}, a, a', \tilde{a}')$ . Under these forms:

- $A^L$  writes  $(a - \delta, a' + \delta, \tilde{a}^{(L)}, \tilde{a}'^{(L)})$  where  $\tilde{a}^{(L)} = (a - \delta, a - \delta, \alpha + \epsilon, 0)$  and  $\tilde{a}'^{(L)} = (a' + \delta, a' + \delta, 0, \alpha' + \epsilon)$ .
- $A_L$  writes  $(a + \delta, a' - \delta, \tilde{a}_{(L)}, \tilde{a}'_{(L)})$  where  $\tilde{a}_{(L)} = (a + \delta, a + \delta, \alpha - \epsilon, 0)$  and  $\tilde{a}'_{(L)} = (a' - \delta, a' - \delta, 0, \alpha' - \epsilon)$ .

**Example:**

Let  $A$  a fuzzy temporal interval representing the period from the early 20 until the end of 20. It is easier to see that  $A$  can writes  $A = (\tilde{\alpha}, \tilde{\alpha}', \alpha, \alpha')$ .

With  $\tilde{\alpha} = (1920, 1920, 3, 0)$  and  $\tilde{\alpha}' = (1930, 1930, 0, 2)$ .

Given that  $\alpha = 3$ ,  $\alpha' = 2$  and  $L = (-3, 3, 1, 1)$ .

Then

$A^L = (1917, 1933, \tilde{\alpha}^{(L)}, \tilde{\alpha}'^{(L)})$  where  $\tilde{\alpha}^{(L)} = (1917, 1917, 4, 0)$  and  $\tilde{\alpha}'^{(L)} = (1933, 1933, 0, 3)$ .

And

$A_L = (1923, 1927, \tilde{\alpha}_{(L)}, \tilde{\alpha}'_{(L)})$  where  $\tilde{\alpha}_{(L)} = (1923, 1923, 2, 0)$  and  $\tilde{\alpha}'_{(L)} = (1927, 1927, 0, 1)$ .

In a similar way, the tolerant counterparts of all Allen relations write:

- $A$  *toler-meets* (L)  $B$  as  $A_L$  before  $B_L$  and  $A^L$  overlaps  $B^L$ .
- $A$  *toler-before* (L)  $B$  as  $A^L$  *toler-meets* (L)  $B^L$ .
- $A$  *toler-overlaps* (L)  $B$  as  $A_L$  *toler-meets*  $B_L$ .
- $A$  *toler-during* (L)  $B$  as  $A^L$  *toler-equals*  $B_L$ .
- $A$  *toler-starts* (L)  $B$  as  $A_L$  *during*  $B^L$  and  $A^L$  *overlaps*  $B_L$ .
- $A$  *toler-finishes* (L)  $B$  as  $A^L$  *overlapped\_by*  $B_L \wedge A_L$  *during*  $B^L$ .
- $A$  *toler-equals* (L)  $B$  as  $A^L$  *contains*  $B_L \wedge A_L$  *during*  $B^L$ .

**3.2. Computation**

In this section, firstly, we recall the comparing indices of two fuzzy intervals proposed by Dubois and Prade in [24].

Let two fuzzy time intervals  $M$  and  $N$  expressed by quadruples of the form  $(m_1, m_2, \alpha_1, \alpha_2)$  and  $(n_1, n_2, \beta_1, \beta_2)$  respectively. Basically, there are four indices to interpret how  $M$  is greater than  $N$  (see Figure 3).

$$d(M \succ N) = \inf_{x,y} \{ \max(1 - \mu_M(x), 1 - \mu_N(y)) : x \leq y \} \quad (5)$$

$$d(M \succ^+ N) = \inf_x \sup_y \{ \max(1 - \mu_M(x), \mu_N(y)) : x \geq y \} \quad (6)$$

$$d(M \succ^- N) = \sup_x \inf_y \{ \max(\mu_M(x), 1 - \mu_N(y)) : x < y \} \quad (7)$$

$$d(M \succ \approx N) = \sup_{x,y} \{ \min(\mu_M(x), \mu_N(y)) : x \geq y \} \quad (8)$$

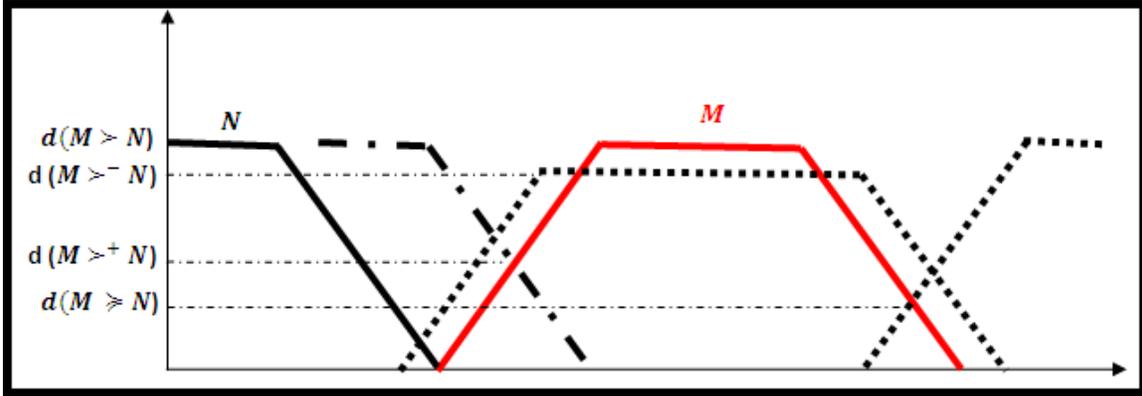


Figure 3. Possibilistic indices of fuzzy intervals comparisons.

Equation (5) expresses the certainty that  $x$  is greater than  $N$ , knowing that  $x$  is  $M$ . This means that  $M$  is necessarily greater than  $N$ . It can be expressed on the basis of a necessity degree of the proposal that  $M$  is strictly greater than  $N$  as follows:

$$\aleph_M(\]N, +\infty)) = 1 - \sup_{x \leq y} \min(\mu_M(x), \mu_N(y)) \quad (9)$$

Equation (9) refers to the degree of inclusion of the fuzzy set  $M$  in  $\]N, +\infty)$ . Knowing that;

$$\aleph_M(\]N, +\infty)) = \text{Ness}(x > N \mid x \text{ is } M)$$

Equation (6) expresses the certainty that  $x$  is greater or equal to  $N$ , knowing that  $x$  is  $M$ . It can be expressed on the basis of a necessity degree of the proposal that  $M$  is greater than or equal to  $N$  as follows:

$$\aleph_M([N, +\infty)) = \inf_x \sup_{y: y \leq x} \max(1 - \mu_M(x), \mu_N(y)) \quad (10)$$

Equation (10) refers to the degree of inclusion of the fuzzy set  $M$  in  $[N, +\infty)$ . Knowing that

$$\aleph_M([N, +\infty)) = \text{Ness}(x \geq N \mid x \text{ is } M)$$

Equation (7) expresses the possibility that  $x$  is greater than  $N$ , knowing that  $x$  is  $M$ . It can be expressed on the basis of a possibility degree of the proposal that  $M$  is strictly greater than  $N$  as follows:

$$\Pi_M(\]N, +\infty)) = \sup_x \inf_{y: y \geq x} \min(\mu_M(x), 1 - \mu_N(y)) \quad (11)$$

Equation (11) refers to the degree of nonemptiness of the fuzzy set  $M \cap \]N, +\infty)$  of numbers strictly greater than  $N$ , given that they are restricted by  $M$ . Knowing that:

$$\Pi_M(\]N, +\infty)) = \text{Poss}(x > N \mid x \text{ is } M)$$

Equation (8) expresses the possibility that  $x$  is greater or equal to  $N$ , knowing that  $x$  is  $M$ . It can be expressed on the basis of a possibility degree of the proposal that  $M$  is greater than or equal to  $N$  as follows:

$$\prod_M([N, +\infty)) = \sup_{x,y : x \geq y} \min(\mu_M(x), \mu_N(y)) \quad (12)$$

Equation (12) refers to the degree of nonemptiness of the fuzzy set  $M \cap [N, +\infty)$  of numbers greater than or equal to  $N$ , given that they are restricted by  $M$ . Knowing that:

$$\prod_M([N, +\infty)) = \text{Poss}(x \geq N \mid x \text{ is } M).$$

Given these degrees of comparison between two fuzzy time intervals, we can use equation (5) redefined in terms of a degree of necessity by the equation (9) to assess the extent to which a fuzzy time interval  $A$  is greater than another fuzzy time interval  $B$  (see Figure 4), denoted by  $d(A > B)$  (with  $A = (a_1, a_2, \alpha_1, \alpha_2)$  and  $B = (b_1, b_2, \beta_1, \beta_2)$ ) as follows;

$$d(A > B) = 1 - \sup_{x \leq y} \min(\mu_A(x), \mu_B(y)) \quad (13)$$

$$= \begin{cases} 1 & \text{if } a_1 - \alpha_1 \geq b_2 + \beta_2 \\ 1 - \rho & \text{if } a_1 - \alpha_1 < b_2 + \beta_2 \text{ and } a_1 > b_2 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{with } \rho = \frac{x - (a_1 - \alpha_1)}{\alpha_1} \text{ and } x = \frac{b_2 \alpha_1 + a_1}{1 + \alpha_1}$$

To illustrate formula (13), let us consider the following example (where  $A = [25/10/2015, 28/10/2015, 1, 1]$  that expresses time around October 25 and October 28):

**Case 1:**  $B = [19/10/2015, 20/10/2015, 2, 2]$

$$a_1 - \alpha_1 > b_2 + \beta_2 \rightarrow 24/07/2015 > 20/07/2015 \text{ then}$$

$$d(A > B) = 1$$

**Case 2:**  $B = [19/10/2015, 24/10/2015, 2, 2]$

$$a_1 - \alpha_1 < b_2 + \beta_2 \rightarrow 24/07/2015 < 26/07/2015 \text{ and}$$

$$a_1 > b_2 \rightarrow 25/07/2015 > 24/07/2015 \text{ then}$$

$$d(A > B) = 1 - \rho = 1 - \frac{b_2 - a_1 + 1 + \alpha_1}{1 + \alpha_1} = 0,5$$

**Case 3:**  $B = [19/10/2015, 26/10/2015, 2, 2]$

$$a_1 - \alpha_1 < b_2 + \beta_2 \rightarrow 24/07/2015 < 27/07/2015 \text{ and}$$

$$a_1 < b_2 \rightarrow 25/10/2015 < 26/10/2015 \text{ then}$$

$$d(A > B) = 0$$

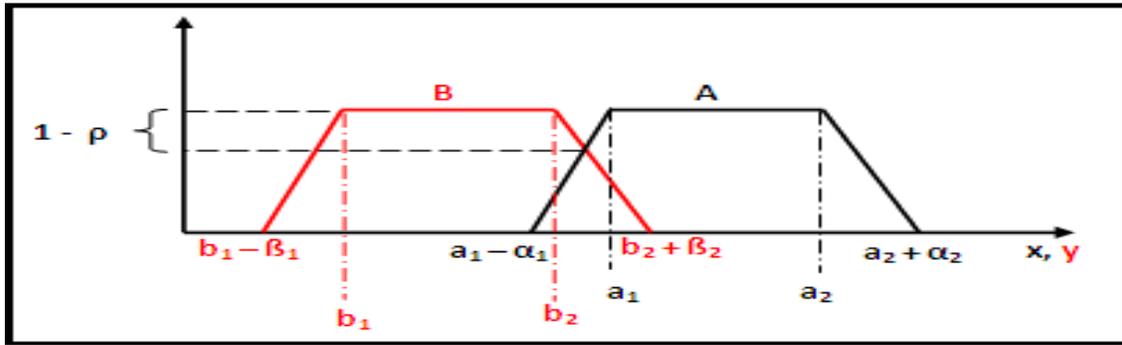


Figure 4. Comparison index  $d(A > B)$ .

Using the formula (13), we define in Table 2 a way to calculate the degree of tolerant Allen relations. We have used this formula because we need to be sure that an interval  $A$  is greater than another  $B$ . This means that  $A$  is necessarily greater than  $B$ .

Table 2. Tolerant Allen Relations (where  $d(\tilde{b}_L > \tilde{a}'_L) = 1 - \sup_{x \leq y} \min(\mu_{\tilde{b}_L}(x), \mu_{\tilde{a}'_L}(y))$ ).

Tolerant Allen Relation	Interpretation	Definition
$A$ toler-meets (L) $B$	$A_L$ before $B_L \wedge A^L$ overlaps $B^L$	$\text{Min}(d(\tilde{b}_L > \tilde{a}'_L), d(\tilde{b}^L > \tilde{a}^L), d(\tilde{a}'^L > \tilde{b}^L), d(\tilde{b}'^L > \tilde{a}^L))$
$A$ toler-before (L) $B$	$A^L$ toler-meets $B^L$	$\text{Min}(d(\tilde{b} > \tilde{a}'), d(\tilde{b}^{2L} > \tilde{a}^{2L}), d(\tilde{b}'^{2L} > \tilde{a}'^{2L}), d(\tilde{a}'^{2L} > \tilde{b}^{2L}))$
$A$ toler-overlaps (L) $B$	$A_L$ toler-meets $B_L$	$\text{Min}(d(\tilde{b}_{2L} > \tilde{a}'_{2L}), d(\tilde{b} > \tilde{a}), d(\tilde{a}' > \tilde{b}), d(\tilde{b}' > \tilde{a}'))$
$A$ toler-during (L) $B$	$A^L$ toler-equals $B_L$	$\text{Min}(d(\tilde{b}_{2L} > \tilde{a}^{2L}), d(\tilde{a}'^{2L} > \tilde{b}'_{2L}), d(\tilde{a} > \tilde{b}), d(\tilde{b}' > \tilde{a}'))$
$A$ toler-starts (L) $B$	$A_L$ during $B^L \wedge A^L$ overlaps $B_L$	$\text{Min}(d(\tilde{a}_L > \tilde{b}^L), d(\tilde{b}'^L > \tilde{a}'_L), d(\tilde{b}_L > \tilde{a}^L), d(\tilde{a}'^L > \tilde{b}_L), d(\tilde{b}'_L > \tilde{a}'^L))$
$A$ toler-finishes (L) $B$	$A^L$ overlapped by $B_L \wedge A_L$ during $B^L$	$\text{Min}(d(\tilde{a}^L > \tilde{b}_L), d(\tilde{b}'_L > \tilde{a}^L), d(\tilde{a}'^L > \tilde{b}'_L), d(\tilde{a}_L > \tilde{b}^L), d(\tilde{b}'^L > \tilde{a}'_L))$

## 4. FUZZ-TIME SYSTEM

### 4.1. Architecture Overview

Here we give an overview of **Fuzz-TIME** system. It includes in fact two main steps and each step contains a set of modules.

Step 1 begins with a temporal gradual query proposed by the user through a GUI of the **Fuzz-TIME** system. The latter requests the user to define a validity interval for each fuzzy temporal specification. Then, *TSQLf* query is generated and sent to a main interpretation module. This module is composed by a set of submodules; each presents an alternative for managing *TSQLf* queries. The result of this module corresponds to an evaluation of an SQL-like query.

The second step firstly proceeds to pass the query by reasoning module. The latter uses the inference machinery of fuzzy Allen relations. Then the request goes through the data management system to select rows that meet the required temporal criteria. Finally, the selected lines pass through the module of our system that calculates the degree of satisfaction of each line with the selection criteria. Then, the selected lines with their degrees are displayed to the user.

### 4.2. Modules Description

In [10] we have proposed an extension, named *TSQLf*, of *SQLf* language [11] by adding the time dimension. *TSQLf* language allows for expressing user queries involving fuzzy criteria on time. It is founded on the fuzzy extension of temporal Allen relations already proposed in [2]. We have implemented this language and developed a first version of the **Fuzz-TIME**<sup>3</sup> system.

In this preliminary version of **Fuzz-TIME**, queries involving fuzzy temporal criteria can be handled where the temporal relations can be defined in a fuzzy way but time are defined only in terms of regular (crisp) intervals. While in real-life applications, time intervals are often described by ill-defined bounds to better capture the vagueness inherent to the available pieces of time information. For this reason, in this work, we have introduced tolerant temporal relations to deal with this issue. Also, we have endowed our **Fuzz-TIME** Tool with the capabilities for handling such novel temporal relations.

First, we present the architecture of **Fuzz-TIME** system. Figure 5 depicts the different modules necessary to processing an *TSQLf* query. In the *Interface* module, the user enters a gradual temporal query using a graphical interface, this latter gives him/her in the first place the possibility of choosing attributes, tables and built fuzzy temporal conditions, and second the system asks the user to identify the validity interval and the tolerance interval for each fuzzy temporal condition (i.e., the fuzzy parameter L).

- If the user provides a validity interval with crisp bounds, then *Fuzzy Allen Relation* module (denoted by FAR) is triggered. More details about this module can be found in [10]. Note that in the case where the *TSQLf* query requires a phase of reasoning, the system calls the module *Reasoning Step*. This module leverages the inference machinery of fuzzy Allen relations developed in the first version.

---

<sup>3</sup> **Fuzz-TIME**: Fuzzy Temporal Information Managing and Exploitation.

- If the user gives a validity interval with fuzzy bounds, then the system automatically computes the eroded and dilated intervals. Thereafter, the system passes the request to the *Query Interpretation* module which transforms the request into a crisp query using the *Tolerant Allen Relations* module (denoted by TAR) using the principle of dilation and erosion operations. The result of this module is an *TSQLf* query to be sent to the management system database in order to select the attributes that meet the fuzzy temporal query criteria.

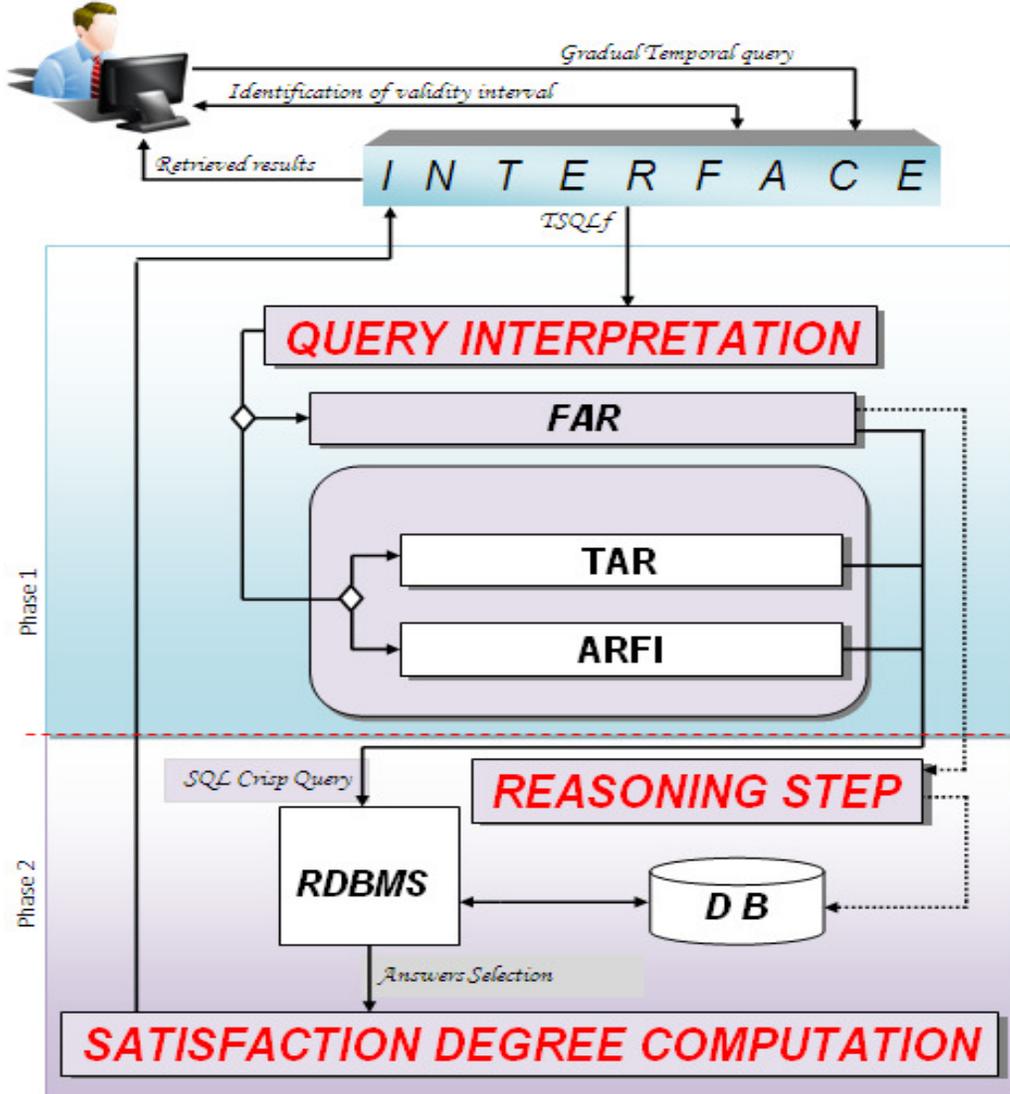


Figure 5. Architecture of **Fuzz-TIME** system.

The last module, *Satisfaction Degree Computation*, ensures the calculation of satisfaction degree of the *TSQLf* query at hand. The returned results are attached with a degree of satisfaction and then displayed on the user interface.

### 4.3. An Illustrative Example

To better explain our proposal we present below an example from Archaeology field. Consider the Archaeology table (see Table 3) presenting the material remains from prehistoric times. The table schema is *Archaeology* (*Code\_Ar*#, *Name\_Ar*, *Location*, *Date\_Discovery*, *VST\_Dc*, *VET\_Dc*, *Date\_Dated*, *VST\_Dd*, *VET\_Dd*). Where *VST\_Dc* means the start validity date of *Date\_Discovery*, *VET\_Dc* means the end validity date of *Date\_Discovery*, *VST\_Dd* means the start validity date of *Date\_Dated* and *VET\_Dd* means the end validity date of *Date\_Dated*.

Table 3. The Archaeology table.

Code_Ar	Name_Ar	Location	Date_Discovery	VST_Dc	VET_Dc	Date_Dated	VST_Dd	VET_Dd
A011	Pyramid of six meters in height	Lima	Recently	10/07/2013	31/07/2013	5000 years ago	2987 BC	2988 BC
A015	Cone rocks	Lake Tiberias	In 2003	05/02/2003	08/08/2003	2050 years ago	37 BC	38 BC
A120	Church	Island of the City	Little ago	20/06/2013	25/07/2013	End 158	08/09/158	18/12/158
A002	Chanel monumental	Narbanne	In 2010	05/03/2010	10/11/2010	2455 years ago	440 BC	442 BC
A020	Hunting weapon	South Africa	Early 2009	03/01/2009	05/04/2009	500000 years ago	497985 BC	497987 BC
A042	Mammoth Skeleton	Seine-et-Marne	Summer 2013	05/06/2013	01/09/2013	Before beginning 159	15/11/189	13/12/189
A075	Sort of mini-dinosaur	Africa	October 2012	02/10/2012	31/10/2012	Before the end of 260	20/09/260	15/12/260
A101	Indian prints	Brazil	Before the end of November 2011	18/11/2011	30/11/2011	More than 3000 years BC	3000 BC	3002 BC
A111	Corps soldiers Allemends	Carspach	Before the end of 2011	12/10/2011	15/12/2011	In 1918	10/07/1918	15/12/1918
A224	Statues banned by the Nazis	Berlin	Beginning in November 2010	01/11/2010	10/11/2010	During the second war mandial	01/09/1939	02/09/1945

### 4.4. Demonstrative Scenarios

We will demonstrate the functioning of our *Fuzz-TIME* system by presenting some particular queries over archaeological databases.

- **Q1:** Show archaeological discoveries that took place during the *3rd quarter* of 2000.

Here the user asked a query with a fuzzy temporal condition (*3rd quarter*) but we can easily interpret the validity interval of the fuzzy temporal condition corresponding to  $[01/07/2000, 30/09/2000]$ . So the solution is a simple SQL query.

```
Select *
From Archaeology A
Where A.VST > 01/07/2000 and A.VET < 30/09/2000;
```

- **Q2:** Show archaeological discoveries that took place just after the Second World War.

The user indicated in this query a fuzzy temporal constraint (Second World War). So we must proceed with the generation of a *TSQLf* request. According to the principle of *TSQLf* query, the

system must ask the user to specify a valid interval for fuzzy temporal constraint proposed by the user. Indeed two cases are possible here:

- i) the user can define a validity interval with two crisp bounds [VST, VET]. For example [01/09/1939, 02/09/1945]

So here we use the first principle of the *TSQLf* approach and the request will be

```
Select *
From Archaeology A
Where A.VET > 02/09/1949;
```

- ii) the user can define a validity interval with two fuzzy bounds [VST, VET]. For example [Before the end of 1939, Just before end of 1945]

So the bounds of validity interval defined by the user are fuzzy, consequently *TSQLf* query should use the principle of Tolerant Allen Relations; the system asks the user to define a fuzzy set  $A = (a, a', \alpha, \alpha')$  for the validity fuzzy interval and a fuzzy set  $L = (-\delta, \delta, \varepsilon, \varepsilon)$ . For example:  $A = (20/09/1939, 10/09/1945, 7, 5)$  and  $L = (-3, 3, 2, 1)$ . Then the system generates automatically two intervals the eroded and dilated ones. Thereafter it generates a *TSQLf* query using Tolerant Allen Relations proposed in section 3. 2.

#### 4.5. Implementation and Interface

We present in this section some user-friendly interfaces that help to make gradual temporal queries based on tolerant Allen relations. It can help users for expressing temporal terms in a fuzzy way. The tool we have developed acts as an JAVA interface with the Oracle DBMS and generates *TSQLf* queries directly executable by calls to functions and PL/SQL bloc stored. The interface is connected to the database so as to store the tables in the field study that incorporates fuzzy temporal aspects. In this way, it is possible to add for each table, which contains a crisp/fuzzy temporal attribute, two specific temporal attributes (VST, VET).

Our first approach proposed in [10] is still functional in our **Fuzz-TIME** system, Figure 6 shows the possibility that a user can enter a query with a temporal fuzzy constraint then he can define the validity interval with two crisp dates.

The extension integrated in **Fuzz-TIME** tool allows, first, to accept the definition of fuzzy temporal condition (Figure 7), and second, the definition of validity interval and tolerance interval for each fuzzy temporal condition. Here the user can introduce fuzzy bounds for each validity intervals (Figure 7).

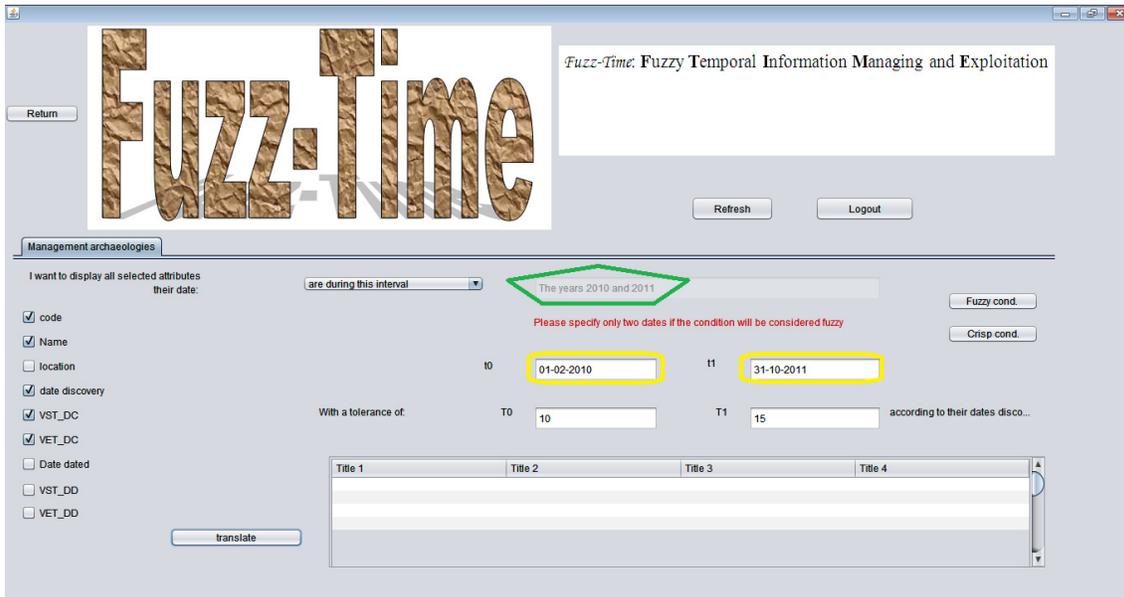


Figure 6. Definition of fuzzy temporal conditions with crisp bounds.

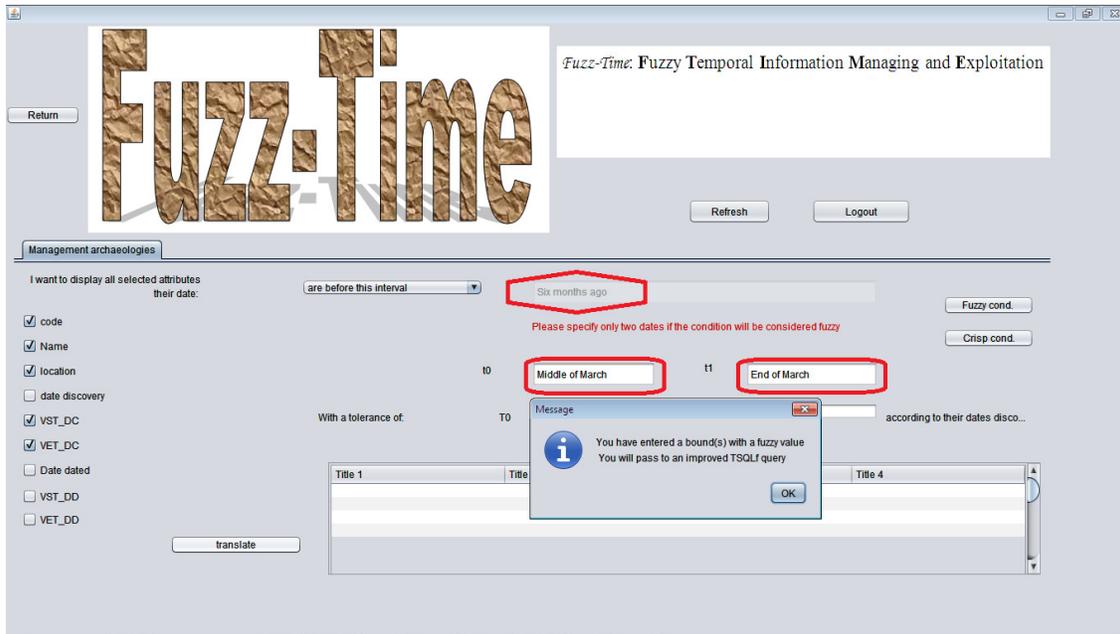


Figure 7. Definition of fuzzy temporal conditions.

After the step above, the tool informs the user that the bounds introduced are fuzzy so it is necessary to define the fuzzy set  $A$  corresponding to a fuzzy temporal condition as well as the fuzzy set  $L$ . The aim is to generate automatically the pair intervals (dilated and eroded ones). Then, the generated TSQLf query is submitted to the database to select rows that match its fuzzy temporal criteria. Finally, a degree of satisfaction is calculated and assigned to each selected line (Figure 8 and Figure 9).

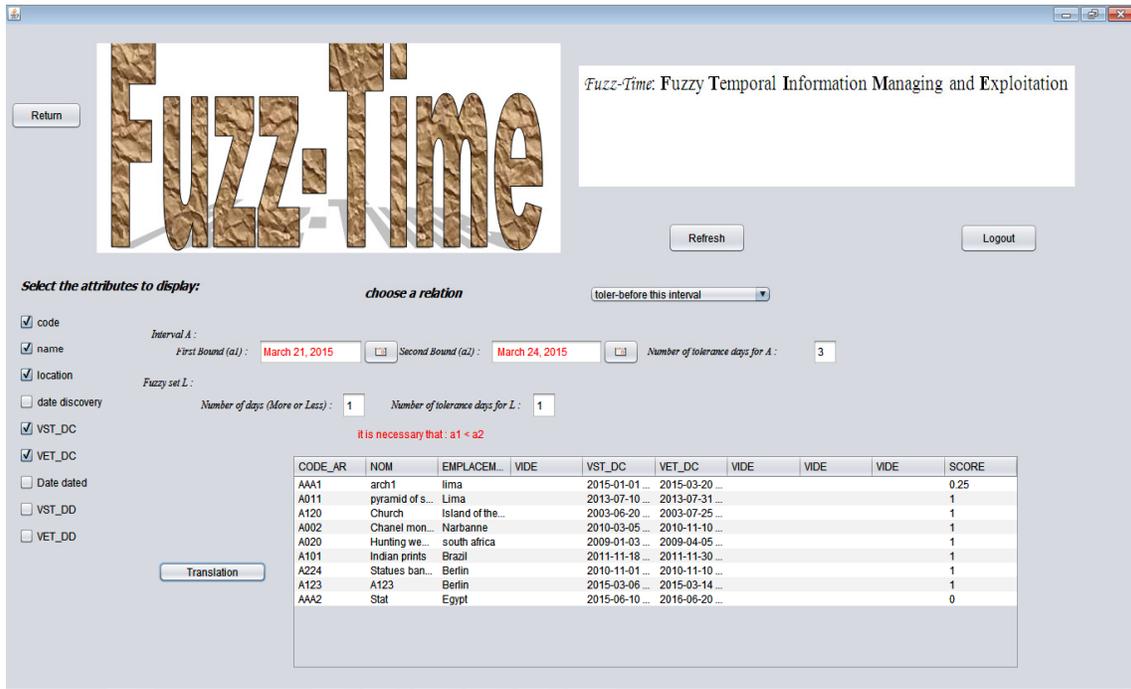


Figure 8. Result of an *TSQlf* query.

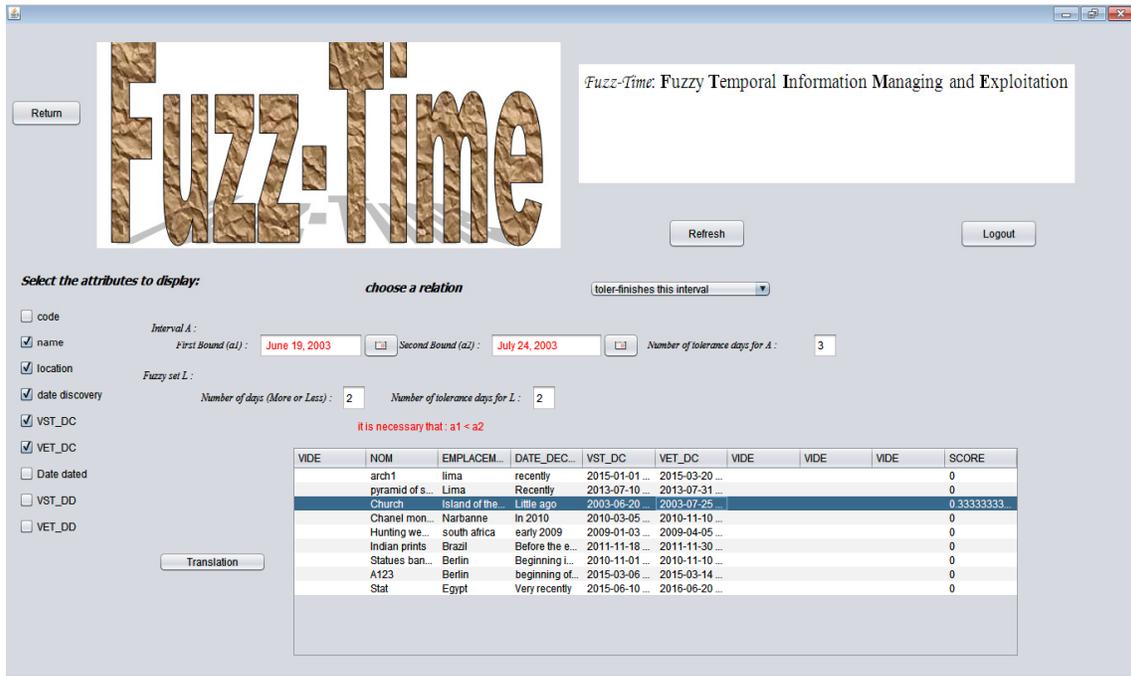


Figure 9. Result of an *TSQlf* query

## 5. CONCLUSION

In this paper, we have defined the principle of tolerance-based of Allen temporal relations to manage time intervals with fuzzy bounds. The key notion of this extension is the dilation and erosion operations defined on fuzzy time intervals. Then we have defined a way for computing such introduced relations by leveraging fuzzy indices comparison. A new version of our **Fuzz-TIME** System is developed. It allows handling temporal queries where time can be expressed in terms of fuzzy intervals.

As for future immediate work, we first plan to establish the complete set of composition rules of the tolerant Allen relations for the purpose of reasoning and inference. Second, we incorporate such reasoning in our **Fuzz-TIME** System. Another line of future research is to investigate the issue of uncertainty in temporal relations in the spirit of [25].

## REFERENCES

- [1] C. Billiet & G. De Tré, (2015) “The Role of Computational Intelligence in Temporal Information Retrieval: A Survey of Imperfect Time in Information Systems”, In Proc. of the 13th International Conference on Flexible Query Answering Systems (FQAS’2015), Vol. 400, pp. 41-756.
- [2] D. Dubois, A. Hadjali, & H. Prade, (2003) “Fuzziness and uncertainty in temporal reasoning”, Journal of Universal Computer Science, Special Issue on Spatial and Temporal Reasoning, Vol. 9, pp. 1168-1194.
- [3] S. Schockaert & M. De Cock, (2008) “Temporal reasoning about fuzzy intervals”, Artificial Intelligence, Vol. 172, pp. 1158-1193.
- [4] S. Badaloni & M. Giacomini, (2006) “The algebra IAfuz: a framework for qualitative fuzzy temporal”, Artificial Intelligence, Vol. 170 (10), pp. 872-902.
- [5] C. Billiet, J. E. Pons, T. Matthe, G. De Tré, & O. P. Capote, (2011) “Bipolar fuzzy querying of temporal databases”, In Proc. of the 9th International Conference on Flexible Query Answering Systems (FQAS’2011), Springer Berlin Heidelberg, Vol. 7022, pp. 60-71.
- [6] F. Pons, C. Billiet, O. Pons, G. De Tré, (2014) “Aspects of dealing with imperfect data in temporal databases”. In: Pivert, O., Zadrozny, S. (eds.) Flexible Approaches in Data, Information and Knowledge Management, Springer, Heidelberg, Vol. 497, pp. 189–220.
- [7] L. Deng, Z. Liang, & Y. Zhang, (2008) “A fuzzy temporal model and query language for fter databases”, 8th International Conference on Intelligent Systems Design and Applications, Vol. 3, pp. 77-82
- [8] C. Tudorie, M. Vlase, C. Nica, & D. Muntraru, (2012) “Modeling fuzzy temporal criteria in database querying”, Artificial Intelligence Applications, Vol. 112, pp. 1-6
- [9] J. Galindo & J.M.Medina, (2001) “Ftsql2 : Fuzzy time in relational databases”, Proc. of the 2nd International Conf. in Fuzzy Logic and Technology, pp. 5-7
- [10] A. Gammoudi, A. Hadjali, & B B. Yaghlane, (2014) “An intelligent flexible querying approach for temporal databases”, Proc. of the 7th international Conference on Intelligent Systems IS’14, Vol. 322, pp. 523-534.
- [11] O. Pivert & P. Bosc, (2012) “Fuzzy Preference Queries to Relational Databases”, Imperial College Press, pp. 330
- [12] J. F. Allen, (1983) “Maintaining knowledge about temporal intervals”, Comm. of the ACM, Vol. 26, pp. 832-843
- [13] D. Dubois & H. Prade, (1988) “Possibility theory”, Plenum Press.
- [14] D. Dubois & H. Prade, (1983) “Inverse operations for fuzzy numbers”, Proc. IFAC Symp. on Fuzzy Info., Knowledge representation and Decision Analysis, pp. 391-395.
- [15] D. Dubois & H. Prade (1989) “Processing fuzzy temporal knowledge”, IEEE Trans. On Systems, Man and Cybernetics, Vol. 19, pp. 729-744.

- [16] H. W. Guesgen, J. Hertzberg, & A. Philpott, (1994) "Towards implementing fuzzy allen relations", Proc. ECAI-94 Workshop on Spatial and Temporal Reasoning, Amsterdam, The Netherlands, pp. 49-55.
- [17] D.Q. Qian & Y.Z. Lu, (1989) "A strategy of problem solving in a fuzzy reasoning network", In Fuzzy Sets and Systems, Vol. 33, pp. 137-154.
- [18] S. Barro, R. Marin, J. Mira, & A. Paton, (1994) "A model and a language for the fuzzy representation and handling of time", In Fuzzy Sets and Systems, Vol. 31. pp. 153-175.
- [19] L. Godo & L. Vila, (1995) "Possibilistic temporal reasoning based on fuzzy temporal constraints", In Proc. of the 14th Inter. Joint Conf. on Artificial Intelligence (IJCAI'95), Montral, Vol. 2, pp. 1916-1922.
- [20] D. Dubois, J. Lang, & H. Prade (1991), "Timed possibilistic logic", In *Fundamentae Informaticae*, Vol. 15, pp. 211-234.
- [21] A. Gammoudi, A. Hadjali, & B B. Yaghlane, (2015) "A Tolerance-Based Semantics of Temporal Relations: First Steps", Proc. of the 7th International Conference on Computational Collective Intelligence Technologies and Applications ICCCI'15, Spring Vol. 9329, pp. 453-462.
- [22] G. Nagypal & B. Motik, (2003) "A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies", On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, LNCS, Vol. 2888, pp. 906-923
- [23] H. Ohlbach (2004), "Relations between fuzzy time intervals". Proc. of the 11th International Symposium on Temporal Representation and Reasoning, pp. 44-51
- [24] D. Dubois & H. Prade, (1983) "Ranking fuzzy numbers in the setting of possibility theory", *Informations Sciences*, Vol. 30, pp. 183-224.
- [25] N. El Hadj Salem, A. Hadjali, B. B. Yaghlane & A. Gammoudi, (2015) "An Evidential Approach to Managing Temporal Uncertainty", Proc. of the 24th French Conference on Fuzzy Logic and its Applications, LFA'15, Poitiers, France, pp. 243-249.

# RELIABILITY EVALUATION OF SOFTWARE ARCHITECTURE STYLES

Gholamreza Shahmohammadi

Department of Information Technology,  
Olum Entazami Amin University, Tehran, Iran  
Shahmohammadi@yahoo.co.uk

## ABSTRACT

*In process of software architecture design, different decisions with system-wide impacts are made. An important decision of design stage is the selection of appropriate software architecture style. Since quantitative impacts of styles on quality attributes have not been studied yet, their application is not systematic. Since Reliability is one of the essential quality requirements of software systems, especially for life critical ones, one of the main criteria in choosing architecture style of these systems is high reliability. The goal of this study is to quantify the impact of architecture styles on software reliability that is desired quality of life critical software. We evaluate styles through reliability block diagram method. First, the reliability equation of each architectural style was computed using of Reliability block diagram approach. Then, reliability rank of architectural styles is computed by setting of the number of effective components in a transaction parameter in reliability equation of architectural styles. The main innovation of this article is quantification of impact of styles on software reliability that is essential for style selection.*

## KEYWORDS

*Software Architecture, Software Architecture Style Evaluation, Reliability block diagram*

## 1. INTRODUCTION

Software systems are increasingly entering consumers' everyday life. To satisfy the consumers' requirements, these systems must demonstrate high reliability and availability. Thus, they must function correctly and without interruption [1]. The software architecture design stage is the first stage of software development in which it is possible to evaluate how well the quality requirements are being met [1]. In the process of architecture design, different decisions are made that have system-wide impact [2]. Architectural decisions made early in the design process are a critical factor in the successful development of system. In particular, the selection of an appropriate software architecture style (SAS) has a significant impact on various system quality attributes [3]. Functionality may be achieved using any of a number of possible structures [4], so SASs are selected based on amount of their support from quality attributes. Styles present models for solving the problem of designing the software architecture in a way that each model describes its components, responsibilities of the components and the way they cooperate [5]. Since quantitative impacts of SASs on quality attributes have not been studied yet [6], their applications

are not systematic [7]. In other words, present use of styles in design is based on intuition of software developers.

Software reliability is widely recognized as one of the most important aspects of software quality [8]. Thus, quantification of impact of SASs on software reliability plays an important role in selecting SASs. This study is a step towards quantification of influence of SASs on quality attributes.

In [9], a method is shown to map an architectural style, expressed formally in an architectural description language, into a relational model that can be checked for various style properties such as consistency. In [3], the impact of a distributed software system's architectural style on the system's energy consumption has been estimated. In [10], a method for specifying the relationship between six SASs and quality attributes such as reliability has been proposed. The relationship between the quality attributes, design principles and some SASs has been specified using a tree-based framework. In [6], impacts of SASs on quality attributes are determined based on the description of style in [11]. [6] and [10] are not able to determine the amount of style support from quality attributes and do not offer quantitative results about their reliability and are not precise.

Different methods for evaluating software reliability are presented [12, 13, and 14]. Considering available information about SASs, only some of those methods like reliability block diagram (RBD) [14] can be used for evaluation of SASs reliability.

In this study, according to the concept of transaction, SASs are evaluated based on reliability block diagram approach and reliability equation of SASs in this approach is determined. Then reliability rank of SASs are determined by setting the number of effective components in a transaction (NECT), and impact of different NECT on SASs rank is investigated. Thus, quantitative impact of SASs on software reliability is determined.

The RBD method presents a block representation of software components and their reliability status, and is a suitable method for determining the reliability of SASs and the estimation of their reliability.

In this work, eight styles: repository (PRS), blackboard (BKB), pipe and filter (P/F), layered (LYD), implicit/invoke (I/I), client/server(C/S), broker (BRK) and object-oriented (OO) are evaluated from reliability viewpoint.

The paper is organized as follows: in section 2 SASs, in section 3 software reliability and their evaluation methods are discussed. In section 4 reliability evaluations of SASs and in section 5 ranking of SASs is described. In section 6 conclusions are explained.

## **2. SASS**

SASs present models for solving the problem of designing the software architecture in a way that each model describes its components, responsibilities of the components and the way they cooperate [11]. Shaw and her colleague [5] introduce seven SASs. Buschmann et al [11] have also described the pattern in different levels.

Since the reliability of SASs are evaluated in this study, regarding effective components in transactions, transaction definition and eight SASs introduced briefly. It should be noted that a transaction is a set of operations that consists a logical unit of the job [15]. Since the evaluation is performed at the level of architectural styles, the transaction is considered in term of effective components in the transaction.

Repository style (RPS). In this style, there are two types of components: a central storage and a set of components that store, retrieve and update information on the repository [5].

Blackboard style (BKB). The components of this style are Blackboard, experts (knowledge resources), and the control. The control component in a loop, checks the blackboard status, evaluates knowledge resource, and activates one of them for the execution [5].

Pipe and filter (P/F). This style is composed of a set of computational components. Each component acts as a filter and has a number of inputs and outputs. The output of each component is the input of the next component [5].

Layered style (LYD). In this style, the emphasis is on different abstraction level in the software. The layered style organized hierarchically. Each layer provides a service for its above layer and uses its lower layer [5].

Implicit Invocation (I/I). Implicit invocation style is an event-driven style based on broadcast concepts and announces the occurrence of the event instead of directly invoking a function. Interested components relate a function to an event. With the occurrence of an event, software invokes all registered functions [5]. Components of this style are: (1) event publishers, (2) components that are interested in events, and (3) dispatcher that invokes interested components in response to an event occurrence.

Client/server(C/S). The components of this style are clients and servers. Clients should be aware of the name and services presented by servers [5].

Broker style (BRK). Client, servers, broker, client side proxy and server side proxy are the components of this style. Broker is responsible for coordinating the relationship between clients and servers. Servers register themselves with the broker, and make their services available to clients through method interfaces. Clients access services of servers by sending requests via the broker. Locating appropriate servers, forwarding the request to it and return the results to the client are the responsibilities of the broker [11].

Object-oriented (OO). In this style, data presentations and the related operations encapsulated in an object. Objects are the components of this style and they interact through invoking the functions [5].

### **3. SOFTWARE RELIABILITY AND THEIR EVALUATION METHODS**

The main objective of software is to offer services desired according to the predetermined quality level. The quality of software has a direct relationship with software architecture. Software often redesigned not because they are functionally deficient, but because they are difficult to maintain, port, or scale[16].

According to IEEE standard 610-12[17] software architecture should provide two types of quality requirements: developmental and operational. Developmental quality Requirements such as maintainability and reusability are important in software development in future. Operational quality requirements such as performance and reliability are important for software users and if the software lacks them, will not be used.

Reliability is a set of sub-characteristics that determines the capability of the software to maintain performance under stated conditions for a stated time period [18]. In other words, reliability indicates a time during which the software is available for use. Reliability sub-characteristics are as follows [18]:

- **Maturity:** the capability of the software product to avoid failures, as a result of faults in the software. The less the frequency of failure, the higher the software maturity will be.
- **Fault tolerance:** the ability to maintain a specified level of performance in case of software fault.
- **Recoverability:** Capability to re-establish the level of performance, Capability to recover the data, and the time and effort needed for it.

The analysis of reliability sub-characteristics indicates the following points:

- The frequency of software failure is directly related to the number of critical components of software architecture, because the more the number of software architecture critical components, the higher the potential possibility of software failure will be.
- Software fault tolerance has a reverse relationship with the number of SAS critical components, because the more the number of SAS critical components, the lower the software fault tolerance will be.
- Recoverability has a reverse relationship with the number of SAS critical components, because principally recoverability is discussed about components by failure of which the performance of software decreases substantially.

Considering the above-mentioned, RBD approach that do reliability evaluation based on software components and their interactions, is suitable method for evaluation of reliability at the level of SAS.

### **3.1. Software Reliability Evaluation Methods**

The software Reliability assessment methods have been classified into quantitative and qualitative methods [1]. There are different types of quantitative methods; some of them are usable before software implementation and some after. Measurement based methods which focuses on the failure of the system are usable before and after software implementation.

RBD method has been presented as a method based on software structure in order to evaluate reliability of software systems. Tripathi et al. [19] have used RBD to estimate reliability of complex software systems from a hierarchy of modules. Leblanc et al. [20] presented a model

based on RBD for indicating real world issues and an algorithm for analysis of this model at the early stage of software development. So this method can be used to evaluate SAS reliability.

### 3.1.1. Reliability Block Diagram (RBD) Method

RBD is a block representation of software components in order to investigate the reliability of components [14]. Since in each architectural style the constituent components the architecture and their interactions are known, RBD method is suitable for evaluating SASs. RBD method, determines SASs reliability based on their structure. SASs reliability prediction is done through investigating the reliability of components. In order to create an RBD the configuration of software architecture components should be determined first. Figure 1 represents a sequential configuration of components. In a sequential configuration, a failure of any component results in failure for the software. The reliability of software with sequential configuration equals the probability that all the components of the architecture of that software succeed. In this case, the reliability of software is computed by Eq. (1). Where  $R_s$  is software reliability,  $X_i$  is event of

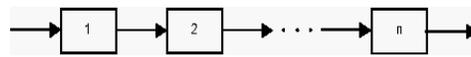


Figure 1. Block diagram of sequential configuration [14]

$$R_s = P(X_1 \cap X_2 \cap \dots \cap X_n) = P(X_1) P(X_2 | X_1) P(X_3 | X_1 X_2) \dots P(X_n | X_1 X_2 \dots X_{n-1}) \quad (1)$$

component  $i$  being operational, and  $P(X_i)$  is the probability that component  $i$  is operational. If components are considered independent then Eq. (1) becomes Eq. (2). In this Eq.,  $R_i$  is the reliability of component  $i$ . In a

$$R_s = \prod_{i=1}^n P(X_i) = \prod_{i=1}^n R_i \quad (2)$$

Sequential configuration, the component with the smallest reliability has the biggest effect on software reliability. As the number of components increases, the software's reliability decreases. Similarly, RBD method can be used for computing software reliability with parallel and sequential-parallel configuration.

## 4. RELIABILITY EVALUATION OF SASS

In section 4.1, RBD method was presented to compute SASs reliability.

Although methods have been offered to predict software components reliability, since the evaluation is done at the architectural styles level, there is no enough information of SASs components. Thus, the reliability of each component is considered  $R_i$ .

### 4.1. Reliability Evaluation of SASs using RBD Method

In this section, SASs will be evaluated using RBD method.

**Repository style (RPS).** In this style, an independent component  $C_i$  interacts with the repository in a transaction. As a result, reliability of style is computed by Eq. (3), where  $R_{rps}$  is the reliability of the repository component, and  $R_i$  is the reliability of each of the independent components.

$$R = R_i R_{rps} \quad (3)$$

**Blackboard style (BKB).** For a transaction in BKB, the control component ( $C_c$ ) checks the blackboard component ( $C_{bb}$ ) status and selects a suitable knowledge resource ( $C_{kr}$ ), then the knowledge resource interacts with the blackboard. Thus, the reliability of style is computed by Eq. (4), where  $R_c$  is the reliability of the control component,  $R_{bb}$  is the reliability of the blackboard, and  $R_{kr}$  is the reliability of the knowledge resource.

$$R = R_c R_{bb} R_{kr} \quad (4)$$

**Pipe and Filter style (P/F).** In this style, in order to perform a transaction all components should be active. Since in each transaction  $m$  components are effective, the reliability of P/F is computed by Eq. (5), where  $R_i$  is the reliability of each filter.

$$R = \prod_{i=1}^m R_i \quad (5)$$

**Layered style (LYD).** In this style in order to perform a transaction, all layers must be active. Thus, reliability of style is also computed by Eq. (5).  $R_i$  is the reliability of each layer.

**Implicit invocation style (I/I).** In this style, a transaction begins with occurrence of an event. Then the event dispatcher component ( $C_d$ ) activates interested component ( $C_i$ ). After the interested component finishes its activity, transaction ends. Thus, the reliability of this style is computed by Eq. (6), where  $R_d$  is the reliability of event dispatcher component and  $R_i$  is the reliability of component interested in the event.

$$R = R_d R_i \quad (6)$$

**Client/Server style (C/S).** In this style, a transaction begins by sending the request of client to the server. After the request is processed by the server that usually needs interaction with repository component, the result is sent to the client. Since the server is a complex component, and consists of several components, the execution of the server is in fact the execution of  $m$  components. Thus, in a transaction,  $m+1$  components, consist of repository and  $m$  components in the server, are each effective. Therefore, the reliability of this style is computed by Eq. (7), where  $R_{rps}$  is the reliability of the repository component and  $R_1 \dots R_m$  are the reliability of server components.

$$R = R_m \dots R_2 R_1 R_{rps} \quad (7)$$

**Broker style (BRK).** In broker style, a transaction begins by sending the request of client. The broker component ( $C_{brk}$ ), client side proxy ( $C_{csp}$ ), server ( $C_s$ ) and repository ( $C_{rps}$ ) are effective in the transaction. Similar to client/server style, the execution of the server is in fact execution of  $m$  components. Thus, in a transaction,  $m+7$  components are effective. so the reliability of this style is computed by Eq. (8), where  $R_{csp}$  is the reliability of client side proxy,  $R_{brk}$  is the reliability of the broker component,  $R_{ssp}$  is the reliability of the server side proxy, and  $R_{rps}$  is the reliability of the repository component.

In server component of client/server and broker styles, a fraction of  $m$  components is effective in transaction. Figure 2 shows the diagram of change in reliability rank of client/server style in terms of the number of server components effective in transaction for  $m = 5$ . According to diagram, by increasing the number of these components the reliability rank of the style is

decreased. This value is considered  $m/2$  based on authors' experience in producing software systems. Software developers determine this parameter based on their former experience in software development.

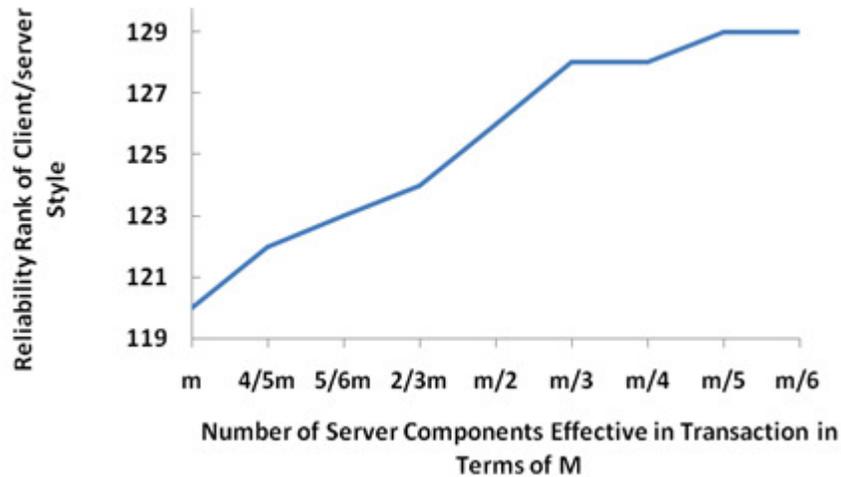


Figure 2. Diagram of change in reliability rank of client/server style in terms of the number of server components effective in transaction for  $m=5$

$$R = R_{csp} R_{brk} R_{ssp} R_m \dots R_2 R_1 R_{tps} \tag{8}$$

**Object-Oriented style (OO).** In this style, the reliability of each use case that consists of  $k$  classes is computed by Eq. (9). An Object-oriented system includes a few use cases, and each use-case has its own execution path. By failure of one use case, the system continues its work with lower throughput. Thus, the configuration of this style is like figure 3, where each execution path indicates a use case. So, the reliability of the style is computed by Eq. (10), where  $p$  is the number of use cases, and  $R_u$  is the reliability of each use case. By substituting  $R_u$  into Eq. (10), the reliability of this style is computed by Eq. (12).

$$R_u = \prod_{i=1}^k R_i \tag{9}$$

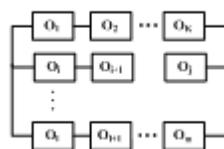


Figure 3. Block Diagram of various transactions

$$R = 1 - \prod_{i=1}^p (1 - R_u) \tag{10}$$

$$R = 1 - \prod_{i=1}^p (1 - \prod_{j=1}^k R_j) \tag{11}$$

In this equation,  $R_u$  reliability of each use case. By substitution of  $R_u$  from Eq. (12), Eq. (11) is obtained. Since some of the classes are common among use cases, the reliability of the style is calculated by Eq. (12) in which  $c$  is the number of common classes in use cases. In object oriented reliability equation,  $c$  indicates the number of common classes in use cases. Investigation

of the reliability of parallel section (  $1 - \prod_{u=1}^p (1 - \prod_{j=1}^k R_j)$  ) with different class number

$$\prod_{i=1}^c R_i (1 - \prod_{u=1}^p (1 - \prod_{j=1}^k R_j)) \tag{12}$$

showed that the reliability of this section is close to 1. So, reliability of the style is equal to that of sequential section (  $\prod_{l=1}^c R_l$  ). So, reliability of this style is calculated by Eq. (13).

$$\prod_{l=1}^c R_l \tag{13}$$

Figure 4 shows the diagram of change in reliability of object oriented style in terms of percent common classes in use cases for  $n_0=15$ . According to diagram, by increasing this percent, reliability of the style is decreased. Based on their experience, the authors considered the percent of common classes in use cases to be 20% of all classes.

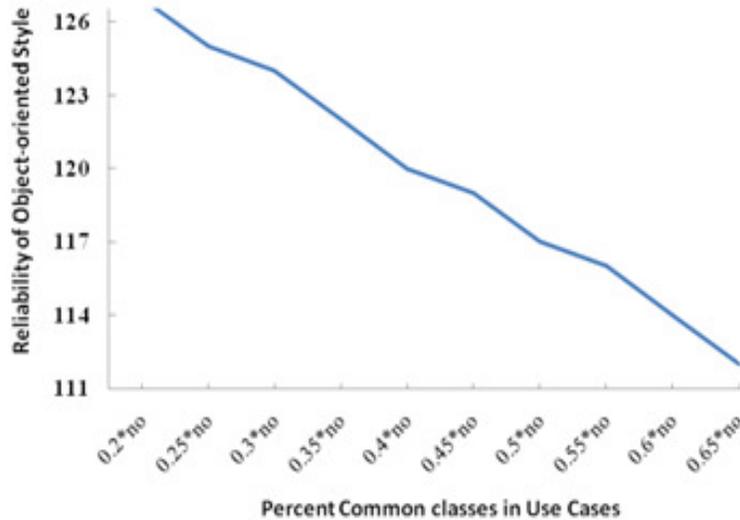


Figure.4. Diagram of change in reliability rank of client/server style in terms of the number of common classes of use cases for  $n_0=15$

Software developers determine common class parameters in use cases and the number of effective server components in transaction based on their former experience in software development.

Table 1 indicates the reliability equation of SASs based on RBD method.

#### 4.2. Reliability Evaluation of Large Systems

In some systems, not only an SAS is used as the main basis for system structuring, but also each of their components may use a specific SAS. Concerning determination of reliability of SASs in this study, it is also possible to evaluate and determine the reliability of such systems too.

Table 1. Reliability equation of SASs based on RBD method

Style	Symbol	Reliability
Repository	RPS	$R_i R_{rps}$
Blackboard	BKB	$R_c R_{bkb} R_{kr}$
Pipe and Filter	P/F	$\prod_{i=1}^m R_i$
Layered	LYD	$\prod_{i=1}^m R_i$
Implicit invocation	I/I	$R_d R_i$
Client/Server	C/S	$R_{m/2} \cdot R_2 \cdot R_1 R_{rps}$
Broker	BRK	$R_{csp} R_{brk} R_{ssp} R_{m/2} \dots R_2 R_1 R_{rps}$
Object-oriented	OO	$\prod_{i=1}^c R_j$

## 5. RANKING OF SASS

In section 4, the reliability of SASs was computed. In the reliability Eq. of most SASs, a parameter 'm' exists which indicates *NECT*. So, in order to investigate the impact of *NECT*, the value of 'm' is considered as 2, 3, 4, 5, 6, and 7. The components of object-oriented style (i.e. classes) are more fine grain than components of other SASs. Thus, with the approximation of three classes in each component in a transaction in average, number of effective classes in a transaction, corresponding to the *NECT* in other SASs, will be considered as 6, 9, 12, 15, 18, and 21 the. We consider the reliability of each component/class 0.98.

For different values of m and based on the reliability Eq. of SASs in tables 1, the reliability of SASs is computed and is shown in tables 2.

Table 2. Reliability of SASs using RBD method

Symbol	m=2	m=3	m=4	m=5	m=6	m=7
	$n_o = 6$	$n_o = 9$	$n_o = 12$	$n_o = 15$	$n_o = 18$	$n_o = 21$
RPS	0.96	0.96	0.96	0.96	0.96	0.96
BKB	0.941	0.941	0.941	0.941	0.941	0.941
P/F	0.96	0.941	0.922	0.904	0.885	0.868
LYD	0.96	0.941	0.922	0.904	0.885	0.868
I/I	0.96	0.96	0.96	0.96	0.96	0.96
C/S	0.96	0.951	0.941	0.932	0.922	0.913
BRK	0.904	0.895	0.886	0.877	0.868	0.859
OO	0.976	0.964	0.953	0.941	0.93	0.919

The reliability rank of SASs has been computed by Eq. (13).

$$y_i = \frac{x_i}{\sum_{i=1}^n x_i} \quad (13)$$

$X_{ij}$  is the reliability value of style in  $i$ -th row and  $j$ -th column. For more clarity, rank values have been shown by coefficient of 1000 in tables 3.

Table 3 shows changes in rank of *SASs* Reliability based on the changes of software size. With increasing software size, rank of some styles such as pipe and filter (*P/F*) and layered (*LYD*) are decreased, rank of some style such as Repository (*RPS*) and implicit invocation (*I/I*) are increased and rank of some style such as Client/server (*C/S*) has not changed considerably

Table 3. Reliability ranks of *SASs* using RBD method

Symbol	m=2	m=3	m=4	m=5	m=6	m=7
	$n_o = 6$	$n_o = 9$	$n_o = 12$	$n_o = 15$	$n_o = 18$	$n_o = 21$
RPS	126	127.1	128.3	129.4	130.6	131.7
BKB	123.4	124.5	125.6	126.7	127.9	129
P/F	126	124.6	123.2	121.9	120.5	119.1
LYD	126	124.6	123.2	121.9	120.5	119.1
I/I	126	127.1	128.3	129.4	130.6	131.7
C/S	126	125.9	125.7	125.6	125.4	125.3
BRK	118.6	118.5	118.4	118.2	118.1	117.9
OO	128.1	127.6	127.3	126.9	126.5	126.1

### 5.1. Analysis of Reliability Rank of Styles According to Reliability of Designed Components

In most *SASs*, the functionalities of some of components are deterministic without considering specific software. Other components are designed regarding the functionalities of the specific software. Thus, the components of each *SAS* are classified into two types: components with specific or constant functionalities, and designed components. For instance, in the broker style, functionalities of broker, client side proxy, and server side proxy components are constant. We consider the reliability of components with constant functionalities as 0.98. As it was mentioned in section 5.3, we consider the reliability of each component at the time of returning from calling of another component as 0.99.

In this section, by changing the reliability of designed components from 0.7 to 0.98, the impact of this change on the reliability of *SASs* for  $m=5$  are studied. According to the graph shown in figure 5, by increasing the reliability of designed components, the rank-difference of the *SASs* reliability decrease. For  $R=0.7$ , the object-oriented style has the lowest and the repository style has the highest reliability rank. By increasing of the reliability of designed components, the trend of reliability rank in the repository, the blackboard and the implicit invocation style is decreasing, while the trend of reliability rank of other *SASs* is increasing.

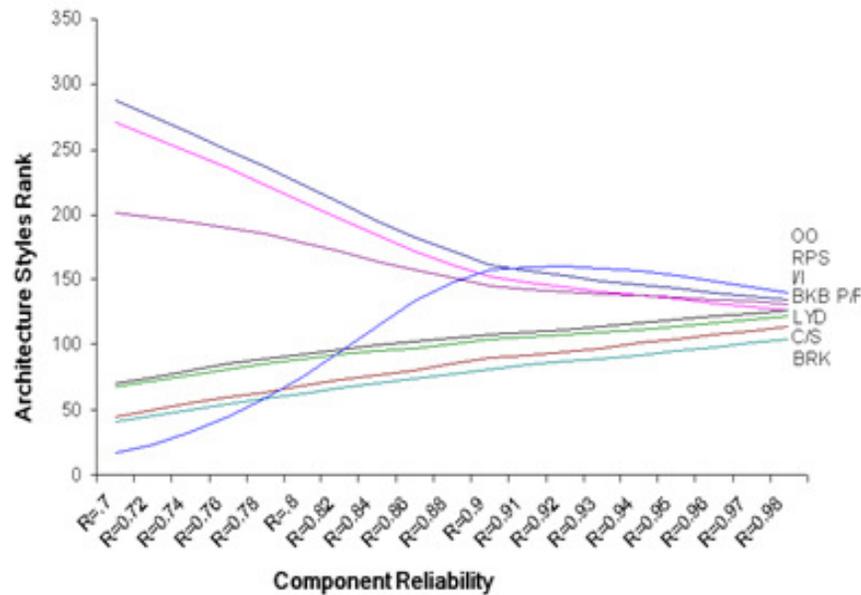


Figure 5. Reliability rank changing trend of SASs based on Change of the reliability of designed components

## 6. CONCLUSION

In this paper, due to the lack of study of quantitative SASs impact on quality attributes specifically reliability; SASs effect on the software reliability was analyzed. First, the reliability equation of each SAS was extracted using of RBD approach. Then, reliability rank of SASs is computed by setting of NECT parameter in reliability equation of SASs in ranges of 2 to 7. Thus, effect of different values of NECT on reliability rank of SASs is determined.

The most important innovation of this paper is quantification of software reliability at the level of SASs and at the stage of SAS selection essential to recommend and select SASs.

Author was evaluated SASs from maintainability viewpoint in [21] and this research extends previous research.

No other similar study was found dealing with evaluation, ranking and comparison of the reliability of SASs. In contrast to [6] and [10], the proposed model is based on architecture evaluation method. The proposed model offers: 1) equations to determine value of reliability of SASs, while [6] and [10] have not done so. In contrast to [6] and [10], the proposed approach offers quantitative results on SASs reliability essential to recommend and select SASs based on recognition of the quantitative effect of SASs on quality attributes. In addition, the proposed approach considered the impact of NECT on rank of SASs reliability. [9] is placed in mathematical model-based evaluation. This method verify features such as consistency and satisfy some features by SASs that are different from desired quality attributes of this paper.

**REFERENCES**

- [1] Immonen A, Niemelä E,(2008) “Survey of reliability and availability prediction methods from the viewpoint of software architecture”, *Journal of Software and Systems Modeling*, Springer, Vol. 7, No.1, pp. 49-65.
- [2] Jansen A G, Bosch J,( 2005) “Software Architecture as a set of Architectural Design Decisions”, 5th IEEE/IFIP Working Conference on Software Architecture, pp. 109–119.
- [3] Seo C, Edwards G, Malek S, Medvidovic N, (2009) “A Framework for Estimating the Impact of a Distributed Software System’s Architectural Style on its Energy Consumption”, 7th Working IEEE/IFIP Conf. on Software Architecture, pp. 277-280.
- [4] Bass L, Clements P, Kazman R, (2003) “Software Architecture in Practice (2nd Edition),” Addison-Wesley.
- [5] Shaw M, Garlan D,( 1996) “Software Architecture: Perspectives Discipline on an Emerging”, Prentice Hall.
- [6] Harrison B, Avgeriou P,( 2007) "Leveraging Architecture Patterns to Satisfy Quality Attributes", 1st European Conf. on Software Architecture, Springer, pp. 263-270.
- [7] Avgeriou P, Zdun U, (2005) “Architectural patterns revisited:a pattern language” ‘Proc. of 10st European Conf. on Pattern Languages of Programs ‘pp.1-39.
- [8] Goseva- Popstojanova K, Trivedi K, (2000)” Failure correlation in software reliability models”. *IEEE Trans. Rel.* 49 1, pp. 37–48.
- [9] Kim J S, Garlan D, (2005) "Analyzing Architectural Styles with alloy", Proc. of the ISSTA 2006 workshop on Role of software architecture for testing and analysis, pp. 70-80, 2006.
- [10] Reza H, Grant E, () ” Quality-Oriented Software Architecture”, Proc. of the IEEE Conf on Information Technology, Coding and Computing, pp. 140- 145.
- [11] Buschmann F, Meunier R, Rohnert H, Sommerlad P, Stal M, (1996) ” Pattern-Oriented Software Architecture- A system of Patterns”. John Wiley & Sons.
- [12] Cheung R C, (1980) “A User-Oriented Software Reliability Model”, *IEEE Transactions on Software Engineering* Vol.6, No 2, pp. 118–125.
- [13] Y. Meng-Lai, C.L.Hyde and L.E. James, (2000) ” A Petri-Net Approach For Early-Stage System-Level Software Reliability Estimation”, Proc. of the IEEE Annual Reliability and Maintainability Symposium, , pp. 100-105.
- [14] Farr W, (1996) Chapter 3 (Software Reliability Modeling Survey), M.R. Lyu (Editor), *Handbook of Software Reliability Engineering*, McGraw-Hill, New York.
- [15] Silberschatz A, Korth H F, Sudarshan S, (1997) ”Database System Concepts”, McGraw-Hill Series in Computer Science.
- [16] Kazman R, Klein M, Barbacci M, Longstaff T, Lipson H, Carriere J, (1998) “ The Architecture Tradeoff Analysis Method”, proc. of the 4th Int. IEEE Conf. on Engineering of Complex Systems. CS Press.,pp.68-78.
- [17] IEEE std 610.12-1990 (n.d.) (1990) IEEE Standard Glossary of Software Engineering Terminology.
- [18] ISO, “ISO 9126-1:2001,( 2001) Software Engineering – Product Quality, Part 1: Quality model”.
- [19] Tripathi R, Mall R, (2005) “Early Stage Software Reliability and Design Assessment“, Proc. of the 12th Asia-Pacific Software Engineering Conference (APSEC’05), Dec.
- [20] Leblanc S P, Roman P A, (2002) “Reliability Estimation of Hierarchical Software Systems”, Proc. of Annual Reliability and Maintainability Symposium.
- [21] Shahmohammadi G.R., (2014) “Evaluation of the Software Architecture Styles From Maintainability Viewpoint”,*int. Journal of computer science & information technology*, Vol. 6, Issue 1.

**AUTHORS**

**Gholamreza Shahmohammadi** received his Ph.D. degree from Tarbiat Modares University (TMU, Tehran, Iran) in 2009 and his M.Sc. degree in Computer Engineering from TMU in 2001. Since 2010, he has been Assistant Professor at the Olum Entezami Amin University (Tehran, Iran). His main research interests are Software Engineering, Software Architecture, Software Metrics, Software Cost Estimation and Software Security.



*INTENTIONAL BLANK*

# A MODEL BASED ON SENTIMENTS ANALYSIS FOR STOCK EXCHANGE PREDICTION - CASE STUDY OF PETR4, PETROBRAS, BRAZIL

Milson L. Lima, Sofiane Labidi, Thiago P. do Nascimento,  
Nadson S. Timbó, Gilberto N. Neto, Marcus Vinicius Lima Batista

Post-graduation Program in Electrical Engineering,  
Federal University of Maranhão, São Luís, Brazil 65080–805  
milsonlima@hotmail.com, soflabidi@gmail.com,  
thiagopinheiro.nascimento@gmail.com, nadsontimbo@gmail.com,  
gilberto.nunes@ifpi.edu.br, marcus\_89lima@hotmail.com

## **ABSTRACT**

*Predicting the behavior of shares in the stock market is a complex problem, that involves variables not always known and can undergo various influences, from the collective emotion to high-profile news. Such volatility, can represent considerable financial losses for investors. In order to anticipate such changes in the market, it has been proposed various mechanisms to try to predict the behavior of an asset in the stock market, based on previously existing information. Such mechanisms include statistical data only, without considering the collective feeling. This article, is going to use natural language processing algorithms (LPN) to determine the collective mood on assets and later with the help of the SVM algorithm to extract patterns in an attempt to predict the active behavior. Nevertheless it is important to note that such approach is not intended to be the main factor in the decision making process, but rather an aid tool, which combined with other information, can provide higher accuracy for the solution of this problem.*

## **KEYWORDS**

*Sentiment Analysis, Twitter, Prediction of Stock Exchanges*

## **1. INTRODUCTION**

Nowadays, with the advancement of Information and Communication Technologies, there had been developed an enabling environment for widespread use of social networks. Such environment is a favorable place for natural exposure of individuals, their desires, preferences and manifestations in its various forms, which makes the result of this process, more natural and close to reality. In cyberspace, each subject is effectively a potential producer of information [1]. We have as a result of this interaction a source of valuable information and yet barely explored. Their use may be the basis for establishing, certain preferences and calculating the mood of individuals. Nowadays, one of the most popular social media is the Twitter, with over 200 million user who share their opinion through tweets.

Bearing in view the large amount of available information, resulting from this process, this paper aims to analyze the content of messages posted on Twitter about certain company in order to establish a relationship between the collective mood of publications and their influence on the price of an asset in the financial market.

The capacity to anticipate the unpredictability of a market, where there are several elements that can influence the value of a stock is desirable on many aspects, especially from a strategic point of view aiming the profit.

## **2. RELATED WORKS**

Frequent exposure of users on social networks, leaves a range legacy of express information in the way that cover a wide variety of topics. And it is on this valuable trace, resulting from the increasingly frantic messages exchange, that many researchers seek to extract valuable information about various fields, making use of various techniques to seek the answers to their questions.

There are several previous works related to sentiment analysis in textual sources using social networks, which goes from predicting the box office revenue for movies [2], consumer opinions about products or services [3], political and policy analysis[4] until disease outbreaks [5].

One who firstly addressed this issue was Pang & Lee [6], which shows in his work an overview of various techniques used for sentiment extraction from the analysis of texts.

Asur and Huberman [7], proposed a model using linear regression, to foresee the box office revenue of a certain movie a few weeks after it is released, it so there would be enough opinions to good analysis. His data set was obtained using Twitter Search API, collected every hour. As research argument they used keywords present in the title of the film, resulting in 2.89 million tweets related to different movies over 3 months. At the end of the task the authors point out the success of their predictions, where it method was more effective than any other used for this domain.

Another highlighted job is from Bollen et al. [8], in it was studied the influence of expressed polarity by Twitter users about particular company and their respective impact on the stock market. Data collection was obtained through Twitter and sentiment analysis was based on the Google-Profile of Mood States (GPOMS) . After analyzing the entire volume of data collected, it was found that variations of collective mood detected, was also observed in the financial market.

## **3. SENTIMENT ANALYSIS**

Sentiment analysis or opinion mining is a branch of mining texts concerned to classify texts not by topic, but by sentiment or opinion contained in a given document. Generally associated with the binary classification between positive and negative sentiments, the term is used more embracing to mean the computational treatment of opinion, sentiment and subjectivity in texts [9].

For Liu [10], The Sentiment Analysis or Opinion Mining is the computational study of opinions, sentiments and emotions expressed in text. The textual information may be classified into two

main types: facts and opinions. The facts are objective expressions about entities, events and their properties. The opinions are generally expressions that describe the sentiment and evaluations of people related to a certain entities, events and their properties.

The sentiment analysis has been one of the most active research areas in the field of Natural Language Processing- NLP and aims to obtain and formalize the opinion and subjective knowledge in unstructured documents (texts) for further analysis within a specific domain [11].

The sentimental analysis has been used in different areas, for example:

- Policy - To measure the popularity of a particular candidate for public office
- Industry - To evaluate the acceptance by consumers to a particular product.
- Stock Exchange - To measure the collective mood on certain asset traded on the stock exchange.

Extract sentiment in textual sources is a complex task, since the natural language processing often comes across expressions which are surrounded by neologisms, irony and other linguistic variations that hinder the correct sentiment extraction.

For our particular case we use the feeling of analysis for the financial area, more specifically in the field Stock Exchange.

## **4. METHODOLOGY**

The topics below demonstrate the methodology applied in this work

### **4.1. The Research Environment**

The environment used for this research was Twitter, a microblogging service, which uses short messages up to 140 characters for information transmission, and can on different platforms [12]. Responsible for about 500 million daily posts [13], it has become a propitious place for researchers and companies seeking for information about the most different subjects through techniques of text mining and natural language processing.

### **4.2. Data Collection**

To access the publications made by users, was used the API from Twitter, provided by the site itself. This is a feature that by user credentials as a developer, allows access to messages through OAUTH5 protocol.

To collect the data, it was used as a criterion messages that contained at least one mention of a word selected, which in our case is represented by the name of the object of our study, Petrobras. The choice of the appropriate term is of utmost importance for the result of the collection of information, since it is the main criteria for the search.

They were collected daily about 3,000 tweets, totaling approximately 40,000 messages between 2015-09-01 and 2015-11-20, in a timetable, between 9:00 AM and 17:00 PM (GMT -02:00) time when the Market was in full operation.

The data resulting from the information gathering process were stored in a database and properly addressed in the pre-processing step in order to remove expressions, unnecessary characters and retweets, aiming not interfere with the review process of individual feeling and collective. Figure 1 illustrates the application process and messages storage.

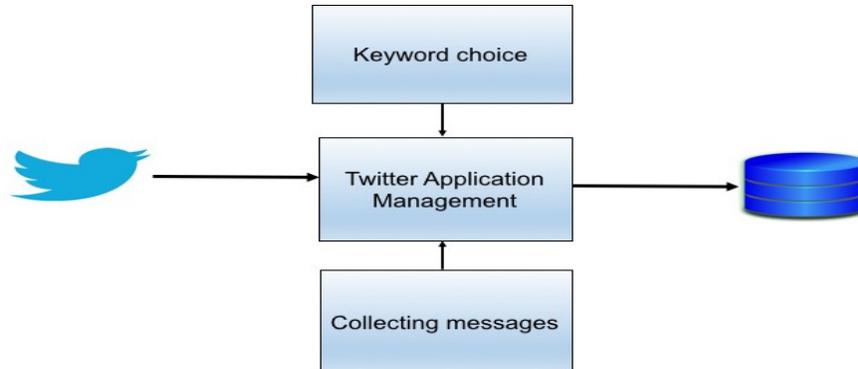


Figure 1. Application Process and storage

### 4.3. Used Tool

After testing some tools available for sentiment analysis, we chose to use the lexicon Sentiment140 [14], that have shown to be more efficient in the classification of messages with twitter features. The table 1 shows the tools used in the testing phase to find the one best suited to this work, as well as some of its features.

Tabel 1. Ferramentas para Análise de Sentimento

Tools	Short Description
SentiWordNet	Lexical dictionary and scores obtained by semi- machine learning approaches
SenticNet	Natural language processing approach for inferring the polarity at semantic level
Sentiment140	API that allows classifying tweets to polarity classes positive, negative and neutral.

In the testing phase were selected 100 tweets randomly within our research field, to be sorted manually and then compared with the results obtained in other tools, as shown in Table 2, where best results are evident when messages were subjected to classification by Sentiment140 lexicon, considering that the polarization done manually in our sample.

Table 2. Hit index to classify tweets

Manually	Senticnet	SentiWordNet	Sentiment140
100%	55%	52%	65%

The Sentiment140 is a lexicon designed specifically to analyze tweets, it corpus was created from a collection of 1.6 million tweets composed of positive and negative emoticons. In it the tweets are labeled in positive or negative according to the respective emoticon. From the auto-labeling

was found that words which occur most frequently in positive or negative tweets, yielding a dictionary with more than 1 million terms, distributed over 62,468 unigrams, bigrams pairs 677,698 and 480,010 [15].

The classification of a tweet sentiment "w" is calculated through the value of its score, as shown below:

$$\text{score}(w) = \text{PMI}(w, \text{positive}) - \text{PMI}(w, \text{negative}) \quad (1)$$

Where, PMI represents pointwise mutual information and receive the default occurrences as positive and negative from the expression, respectively. A positive score indicates association with the positive sentiment, while a negative score indicates association with the negative sentiment.

Like the majority of the methods and techniques available for this purpose contents is available only in the English language.

#### 4.4. Analyzing Sentiment

In order to obtain the daily collective sentiment, which are commented on Twitter about in our dominion that were randomly selected 1,000 tweets per day in Portuguese from Brazil. After the preprocessing steps and data transformation data, the result was arranged as shown in Table 3, which demonstrate a period of the fragment under analysis (2015-10-08 to 2015-10-26), where there is clearly a prevalence of "negative" mood over "positive", the latter won only two instances, one on 2015-10-09 and another on 2015-10-26.

Table 3 – Daily Collective Sentiment

DAY	MESSAGES (Tweet's)			
	POSITIVE	NEUTRAL	NEGATIVE	SENTIMENT
2015-10-08	366	99	535	
2015-10-09	499	74	425	
2015-10-12	355	88	443	
2015-10-13	408	121	471	
2015-10-14	344	103	553	
2015-10-15	398	112	602	
2015-10-16	323	96	581	
2015-10-19	313	80	607	
2015-10-20	406	79	515	
2015-10-21	248	63	687	
2015-10-22	290	123	587	
2015-10-23	403	149	448	
2015-10-26	521	93	386	

The highest occurrence of "negative" mood especially in larger quantities our sample, can be attributed to recent episodes of financial scandals involving Petrobras company which is the subject of our study.

#### 4.5. Machine Learning and Sentiment Analysis

Within the process of sentiment analysis and prediction, a major challenge is to find an efficient way to classify the texts for analysis. The classification process consists to find, through machine learning, a function that expresses the best possible way, data classes involved in the field and thereby, making automatic the classification process for new instances, with reference to the model for which it was trained.

Traditional machine learning approaches to text classification, as Naive Bayes (NB), Maximum Entropy (ME) and Support Vector Machines (SVM), are quite effective when applied to sentiment analysis problem [16].

In this article we are going to use a data set previously labeled for the supervised training, these data were divided into two groups, one for training, where we are going to use 70% of the instances and another to test with 30%.

The tool used for the knowledge process discovery was the WEKA and attributes chosen for training and testing were a historical series within the containing period in question: opening price, minimum price, maximum price, closing price and closing situation (high / low). We use two algorithms classification widely used for this purpose, the Support Vector Machine (SVM) and Naive Bayes, in search for the best accuracy for this first stage of our tests. Table 4 shows the results obtained for each.

Table 4 - Comparison of Naive Bayes and SVM

Algorithm	Correctly Classified Instances
Naive Bayes	35.29%
SVM	47.05%

In a second step, and now using only the SVM, for been shown better results in the first stage, it was inserted the attribute "sentiment" that can assume two values: positive or negative, at the same set of data and maintaining the same proportionality between training and testing. Table 5 shows the result of the insertion of this attribute for the accuracy of the chosen classifier.

Table 5 - SVM after inserting the attribute "sentiment"

Algorithm	Correctly Classified Instances
SVM	88.23%

#### **4.6. Results Obtained**

It is noticeable that the insertion of the attribute "sentiment" into the set of data processed by the classifier, resulted in a significant gain in the correct classification of instances using the SVM algorithm.

One of the factors that certainly contributed for this result was not high number of instances that was used to train our classifier, specifically 56. In future work we will repeat the same tests for a much longer period and to seek to ratify the relationship between the collective mood and the financial market.

### **5. CONCLUSIONS**

In this investigation when using Twitter data, it was found that to use them satisfactorily for this purpose, it is necessary a great process to treat the information, since the messages exchanged on this platform are rich in ironies, neologisms and slang which greatly complicates the analysis of the feeling contained in tweets.

Another factor worth mentioning is that not always the collective mood is the true sentiment is about a particular asset, we take as an example the case study in the Brazilian company Petrobras. It is a state-owned joint stock, which in its present time has become entangled in a series of political scandals and misuse of public money. Much of comments where it appears the company name, has an essentially political connotation, that is, there is a shift of focus in the comments, many users write messages containing insults to politicians, for example, which will certainly happen in a scenario where politics the country were bad and the company had its valued stocks, in other words, companies with these characteristics may not be sensitive to the public mood, simply because they have their image associated with the state. The present moment the company with its shares rising devaluation also coincides with the unfavorable political moment which may explain the results presented in the study.

which demonstrate that the mood predominantly negative in the days study, followed by a devaluation of the shares.

Another approach that can be developed from this work is based on the same methodology applied, expand it to companies of different sizes as small caps, which are of low market value companies and study their sensitivity to the collective mood, in other words, it is to evaluate whether the company's capital size is a factor to be considered for this type of study.

Despite the studies focused on sentiment analysis have evolved significantly in recent years, the tools resulted from this process should be seen as something complementary in the decision making process, given the complexity of extracting exactly sentiment textual sources in natural language.

**REFERENCES**

- [1] LEMOS, Lúcia. O poder do discurso na cultura digital: o caso Twitter. *Revista de Estudos e Pesquisas em Linguagem e Mídia*, São Paulo, v.4. n.1. Janeiro-Abril de 2008.
- [2] S. Asur and A. Huberman. Predicting the Future with Social Media. *CoRR*, abs/1003.5699. 2010.
- [3] Eirinaki, M., Pisal, S., and Singh, J. Feature-based opinion mining and ranking. *Journal of Computer and System Sciences* 78, 4 (July 2012), 1175–1184.
- [4] Fang, Y., Si, L., Somasundaram, N., Yu, Z.: Mining contrastive opinions on political texts using cross-perspective topic model. In: *Proceedings of the fifth ACM international conference on Web search and data mining - WSDM'12*. p. 63. ACM Press, New York, USA (2012).
- [5] St Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks? *BMJ*, 344, 1-3.
- [6] Pang, B., & Lee, L. (2008). *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval. doi:10.1561/1500000011
- [7] S. Asur and A. Huberman. Predicting the Future with Social Media. *CoRR*, abs/1003.5699. 2010.
- [8] BOLLEN, J., MAO, H., AND ZENG, X. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.
- [9] PANG, Bo;LEE, Lilian. (2008). *Opinion Mining and Sentiment Analysis*. Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, June 2008.
- [10] Liu, B. (2010). *Sentiment Analysis and Subjectivity*. In Nitin Indurkha & F. J. Damerau (Eds.), *Handbook of Natural Language Processing*, Second Edition. Boca Raton, FL: CRC Press, Taylor and Francis Group.
- [11] Liu, B.: *Sentiment Analysis and Opinion Mining*. *Synthesis Lectures on Human Language Technologies* 5(1), 1–167 (May 2012).
- [12] Stevens, Vance (2008). “Trial by Twitter: The Rise and Slide of the Year’s Most Viral Microblogging Platform”. *TESLEJ: Teaching English as a Second or Foreign Language*, Vol. 12, N. 1, 2008.
- [13] <https://blog.twitter.com/2014/the-2014-yearontwitter>, accessed: (2015-09-12).
- [14] Alec Go, Richa Bhayani, and Lei Huang. 2009. *Twitter Sentiment Classification using Distant Supervision*. In *Final Projects from CS224N for Spring 2008/2009 at The Stanford Natural Language Processing Group*.224N for Spring 2008/2009 at The Stanford Natural Language Processing Group.
- [15] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the International Workshop on Semantic Evaluation, SemEval '13*, Atlanta, Georgia, USA, June.
- [16] Read, J (2005). Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification. *Proceedings of the ACL Student Research Workshop*. 43-48.

## AUTHORS

### Milson Louseiro Lima

Postgraduate Diploma in ANALYSIS AND SYSTEMS PROJECT (UFMA, 2007), graduated in Economic Sciences from the Federal University of Maranhão, Brazil-UFMA, (UFMA,2006). Is ERP developer and mobile devices, since 1998. He is currently a Master's student Electrical Engineering course for Computer Science (UFMA, 2014), working in the Intelligent Systems Laboratory at the Federal University of Maranhão(LSI/UFMA).



### Sofiane Labidi

Bachelor's at Ciência da Computação from Institut Supérieur Scientifique (1990), master's at Ciência da Computação from Université de Nice Sophia Antipolis Centre National de Recherches Scientifiques (1991) and doctorate at Ciência da Computação from Institut National de Recherche en Informatique et Automatique (1995). He is currently full professor at Universidade Federal do Maranhão. Has experience in Computer Science, acting on the following areas: knowledge management, multi-agent systems, educational technologies, agents, artificial intelligence and business proces modelling.



### Thiago Pinheiro do Nascimento

Bachelor's at Ciência da Computação from Faculdade de Ciências Humanas, Saúde, Exatas e Jurídicas de Teresina (2012) and master's at Electric Engineering from Universidade Federal do Maranhão (2015). Has experience in Computer Science, acting on the following subjects: frameworks, software engineering component-based, service-oriented architecture, software reuse and web development.



### Nadson Silva Timbó

Has graduation at Ciencia da Computação by Universidade Federal do Maranhão (2013). Currently is of Universidade Ceuma. Has experience in the area of Computer Science. management, multi-agent systems, educational technologies, agents, artificial intelligence and business proces modelling.



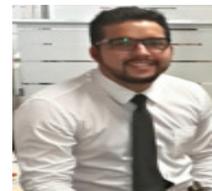
### Gilberto Nunes Neto

bachelor's at Licenciatura Plena em Computação from Universidade Estadual do Piauí (2006). Has experience in Computer Science, focusing on Computer Science



### Marcus Vinicius Lima Batista

Bachelor in Information Systems from the University CEUMA. Researcher at the Center of Research and Extension in Technology and Information Systems (NUSTI). Management studies in T.I; Research online e-Government



*INTENTIONAL BLANK*

# OPTIMAL BEAM STEERING ANGLES OF A SENSOR ARRAY FOR A MULTIPLE SOURCE SCENARIO

Sanghyouk Choi<sup>1</sup>, Joohwan Chun<sup>1</sup>, Inchan Paek<sup>2</sup> and Jonghun Jang<sup>3</sup>

<sup>1</sup>Department of Electrical Engineering, KAIST, Daejeon, Korea  
shchoi@sclab.kaist.ac.kr

<sup>2</sup>PGM Image Sensor Centre, Hanwha Thales, Kyunggi-do, Korea  
ic.paek@hanwha.com

<sup>3</sup>Agency for Defense Development, Daejeon, Korea

## ABSTRACT

*We present the gradient and Hessian of the trace of the multivariate Cramér-Rao bound (CRB) formula for unknown impinging angles of plane waves with non-unitary beamspace measurements. These gradient and Hessian can be used to find the optimal beamspace transformation matrix, i.e., the optimum beamsteering angles, using the Newton-Raphson iteration. These trace formulas are particularly useful to deal with the multiple source scenario. We also show the mean squared error (MSE) performance gain of the optimally steered beamspace measurements compared with the usual DFT steered measurements, when the angle of arrivals (AOAs) are estimated with stochastic maximum likelihood (SMLE) algorithm.*

## KEYWORDS

*Cramér-Rao bound, beamspace transformation, correlated noise, DFT beams, subarray formation*

## 1. INTRODUCTION

The angle of arrival (AOA) estimation problem arises a wide variety of applications dealing with electromagnetic or sound waves. Therefore, AOA estimation has been one of the most active research topics for the past several decades, and various algorithms such as the interferometry, monopulse, MUSIC, ESPRIT and maximum likelihood estimation (MLE) have been devised. A related problem of theoretical importance is to get the Cramér-Rao bound (CRB), i.e., a lower bound of the mean squared error (MSE) for the AOA estimation, which also has been studied by many researchers. However, what has been missing in this direction of research is the investigation of the effect of controllable parameters on the CRB. In this paper, we derive the optimal beam directions that minimizes the CRB for the AOA estimation problem in the presence of multiple impinging plane waves. We also show that the resulting optimal directions give a lower MSE through Monte-Carlo simulation.

The angle of arrival (AOA) estimation may be carried out either in the sensor element space or in the beam space. Although accurate angle estimation requires an array with a large number of

sensors in general, direct utilization of all sensors is impractical. The reason is that manipulation of full multi-channel digital data incurs high computational burden, let alone the high hardware cost for down-conversion and digitization of every sensor measurement signal. To alleviate the difficulty, a dimension reduction matrix  $B \in \mathbb{X}^{M \times K}$ ,  $K < M$ , may be used to transform an element space (ES) measurement vector  $z \in \mathbb{X}^{M \times 1}$  to a beamspace (BS) measurement vector  $B^H z \in \mathbb{X}^{K \times 1}$ , where superscript  $H$  denotes the conjugate transpose. The BS transformation matrix,  $B$  may be designed under different criteria, for example, to cover a given spatial sector [5], to maximize the signal to noise ratio (SNR) in the sector [11], to minimize the interfering power [10], to minimize the Cramér-Rao bound (CRB) [1], or simply to ease the implementation by employing the discrete Fourier transformation (DFT) [13], [2], [12]. Apart from the DFT-based transformation, the above-mentioned  $B$  will have arbitrary complex-numbered elements, and therefore, they will incur higher hardware cost than the case with unit-modulus complex numbers, if  $B$  is to be implemented with analog parts.

If we design  $B$  with unit-modulus elements, it admits a simpler implementation using phase-shifters only. In other words, we wish to steer the beams tactfully, in a different manner from the usual orthogonal DFT beams, so that the CRB for the angle estimation is the minimum. A block diagram of the proposed transformation is shown in Fig 1.

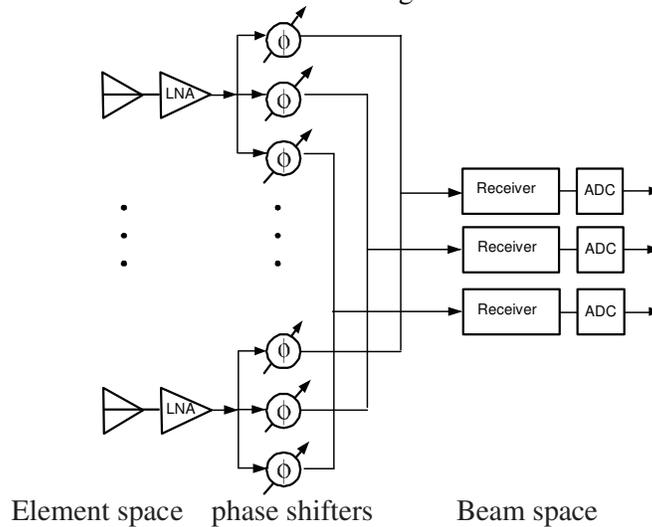


Figure1: Beamspace transformation with sphase-shifters only (K=3)

Now the CRB formula in [9] cannot be used to find the optimum steering angles because  $B$  is not unitary and therefore the measurement noise is spatially correlated. Furthermore, the pre-whitening technique [4, 6, 7] cannot be used because the pre-whitening matrix will not be an explicit parametric function of the steering angles, and therefore does not admit explicit differentiations.

## 2. REVIEW OF THE CRB FOR BEAM SPACE MEASUREMENT

Consider the measurement vector  $z \in \mathbb{X}^{M \times 1}$  of an array with  $M$  sensors,

$$z = As + n, \quad A = [a(\theta_1), \Lambda, a(\theta_N)], \quad s = [s_1, \Lambda, s_N]^T \quad (1)$$

where the  $n$ th column  $a(\theta_n)$  of  $A \in \mathbb{X}^{M \times N}$  is the array response vector for the  $n$ th plane-wave signal  $s_n$  impinging at the angle  $\theta_n$ , superscript  $T$  denotes the transpose, and  $n : CN(0, \sigma_n^2 I)$  denotes the spatially and temporally uncorrelated measurement noise vector. Then, the stochastic CRB of  $\theta = [\theta_1, \Lambda, \theta_N]^T$  is given by [9]

$$CRB(\theta) = \frac{\sigma_n^2}{2L} \left\{ \text{Re} \left[ \left( D^H \Pi_A^\perp D \right) \varepsilon \left( PA^H R^{-1} AP \right)^T \right] \right\}^{-1}, \quad (2)$$

where  $R = Ezz^H = APA^H + \sigma_n^2 I$ ,  $P = Ess^H$ ,  $\Pi_A^\perp = I - A(A^H A)^{-1} A^H$ ,  $D = [d_1, \Lambda, d_n]$ ,  $d_n = \frac{da(\theta_n)}{d\theta_n}$ ,  $L$  is the number of snapshots,  $\text{Re}(\cdot)$  is the real-part operator, and  $\varepsilon$  denotes the Hadamard-Schur product.

Let  $B = [a(\theta_{b1}), \Lambda, a(\theta_{bk})] \in \mathbb{X}^{M \times K}$  denote a BS transformation matrix, towards a steering angle  $\theta_b = [\theta_{b1}, \Lambda, \theta_{bk}]^T$ . Then the resulting dimension reduced measurement will be

$$z_b = B^H As + B^H n. \quad (3)$$

In general, the noise vector  $B^H n$  is correlated, which may be pre-whitened as  $z_b = UB^H As + UB^H n$ , where  $U$  is a Hermitian square root matrix of  $(B^H B)^{-1}$ . Now the stochastic BS CRB can be expressed as

$$CRB_B(\theta, \theta_b) = \frac{\sigma_n^2}{2L} \left\{ \text{Re} \left( F \varepsilon G^T \right) \right\}^{-1}, \quad (4)$$

where  $F = D^H BU^H \Pi_{UB^H A}^\perp UB^H D$ ,  $G = PA^H BU^H R^{-1} UB^H AP$  and

$R = UB^H APA^H BU^H + \sigma_n^2 I$ . However,  $U$  is a function of  $\theta_{bk}$  whose explicit form is not available, and therefore, we cannot minimize (4) with respect to  $\theta_{bk}$ . The following lemma [12] gives a direct CRB formula for (3), not resorting to pre-whitening.

**Lemma 1** For the BS measurement model in (3), the CRB for  $\theta$  is given as

$$CRB_B(\theta, \theta_b) = \frac{1}{2L} \left\{ \text{Re} \left( F \varepsilon G^T \right) \right\}^{-1} \quad (5)$$

where  $F = D^H \Pi_{\Pi_B^H A}^\perp BR^{-1} B^H D$ ,  $G = \left( PA^H BR^{-1} B^H AP \right)^T$  and  $R = B^H APA^H B + \sigma_n^2 B^H B$ .

### 3. GRADIENT AND HESSIAN OF THE TRACE OF THE CRB

Assuming that multiple sources impinge on an array of sensors at angles,  $\boldsymbol{\theta} = [\theta_1, \Lambda, \theta_{N_s}]^T$ , we are interested in finding the optimal steering angles  $\boldsymbol{\theta}_b$  for the BS transformation matrix  $B = [a(\theta_{b1}), \Lambda, a(\theta_{bk})]$  such that

$$\hat{\boldsymbol{\theta}}_b = \underset{\boldsymbol{\theta}_b}{\operatorname{argmin}} \operatorname{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)]. \quad (6)$$

Since  $\operatorname{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)]$  is a nonlinear function of  $\boldsymbol{\theta}_b$ , the minimization needs to be carried out numerically.

We shall use the Newton-Raphson iteration which needs the gradient  $\nabla \operatorname{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)]$  and Hessian  $H(\operatorname{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)])$  of  $\operatorname{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)]$ .

They are provided in the following lemmas.

**Lemma 2** *The  $i$ th element of the gradient vector,  $\nabla \operatorname{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)]$  is given by*

$$\frac{\partial \operatorname{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)]}{\partial \theta_{bi}} = -\frac{1}{2N} \operatorname{tr} \left[ \operatorname{Re}(F \boldsymbol{\varepsilon} G)^{-1} \cdot \operatorname{Re}(U_i \boldsymbol{\varepsilon} G + F \boldsymbol{\varepsilon} V_i) \cdot \operatorname{Re}(F \boldsymbol{\varepsilon} G)^{-1} \right] \quad (7)$$

where

$$F = (D^H \Pi_{\Pi_B A}^\perp B R^{-1} B^H D)$$

$$G = (P A^H B R^{-1} B^H A P)^T$$

$$U_i = \frac{\partial F}{\partial \theta_{bi}} = (D^H H_i B R^{-1} B^H D) + (D^H \Pi_{\Pi_B A}^\perp C_i R^{-1} B^H D) \\ + (D^H \Pi_{\Pi_B A}^\perp B R^{-1} C_i^H D) + (D^H \Pi_{\Pi_B A}^\perp B L_i B^H D)$$

$$V_i = \frac{\partial G}{\partial \theta_{bi}} = (P A^H C_i R^{-1} B^H A P)^T + (P A^H B R^{-1} C_i^H A P)^T + (P A^H B L_i B^H A P)^T$$

$$C_i = \frac{\partial B}{\partial \theta_i}$$

$$L_i = \frac{\partial R^{-1}}{\partial \theta_{bi}} = -R^{-1} G_i R^{-1}$$

$$H_i = \frac{\partial \Pi_{\Pi_B A}^\perp}{\partial \theta_{bi}} = \frac{\partial}{\partial \theta_{bi}} \left( I - \Pi_B A (A^H \Pi_B A)^{-1} A^H \Pi_B \right) \\ = -\frac{\partial}{\partial \theta_{bi}} \Pi_B A (A^H \Pi_B A)^{-1} A^H \Pi_B$$

$$\begin{aligned}
&= -O_i A (A^H \Pi_B A)^{-1} A^H \Pi_B - \Pi_B A (A^H \Pi_B A)^{-1} A^H O_i - \Pi_B A P_i A^H \Pi_B \\
P_i &= \frac{\partial}{\partial \theta_{bi}} (A^H \Pi_B A)^{-1} = -(A^H \Pi_B A)^{-1} (A^H O_i A) (A^H \Pi_B A)^{-1} \\
O_i &= \frac{\partial}{\partial \theta_{bi}} \Pi_B = C_i (B^H B)^{-1} B^H + B Q_i B^H + B (B^H B)^{-1} C_i^H \\
Q_i &= \frac{\partial}{\partial \theta_{bi}} (B^H B)^{-1} = -(B^H B)^{-1} (C_i^H B + B^H C_i) (B^H B)^{-1} \\
G_i &= \frac{\partial R}{\partial \theta_{bi}} = C_i^H A P A^H B + B^H A P A^H C_i + \sigma_n^2 C_i^H B + B^H C_i
\end{aligned} \tag{8}$$

*Proof.* The proof of this lemma is omitted because it is tedious but straightforward, utilizing matrix differentiation rules.

**Lemma 3** The  $(i, j)$  th element of the Hessian matrix  $H(\text{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)])$  is given by

$$\begin{aligned}
\frac{\partial \text{tr}[CRB(\boldsymbol{\theta}, \boldsymbol{\theta}_b)]}{\partial \theta_{bi} \partial \theta_{bj}} &= -\frac{1}{2N} \text{tr} \left[ \text{Re} \left( \frac{\partial Q}{\partial \theta_{bj}} \right)^{-1} \cdot \text{Re} \left( \frac{\partial Q}{\partial \theta_{bi}} \right) \cdot \text{Re}(Q)^{-1} \right. \\
&\quad \left. + \text{Re}(Q)^{-1} \cdot \text{Re} \left( \frac{\partial Q}{\partial \theta_{bi} \partial \theta_{bj}} \right) \cdot \text{Re}(Q)^{-1} + \text{Re}(Q)^{-1} \cdot \text{Re} \left( \frac{\partial Q}{\partial \theta_{bi}} \right) \cdot \text{Re} \left( \frac{\partial Q}{\partial \theta_{bj}} \right)^{-1} \right]
\end{aligned} \tag{9}$$

where  $Q = F \boldsymbol{\varepsilon} G$  and

$$\begin{aligned}
\text{Re} \left( \frac{\partial Q}{\partial \theta_{bj}} \right)^{-1} &= -\text{Re}(Q)^{-1} \cdot \text{Re} \left( \frac{\partial Q}{\partial \theta_{bj}} \right) \cdot \text{Re}(Q)^{-1} \\
\text{Re} \left( \frac{\partial Q}{\partial \theta_{bi}} \right) &= \text{Re}(U_i \boldsymbol{\varepsilon} G + F \boldsymbol{\varepsilon} V_i) \\
\text{Re} \left( \frac{\partial Q}{\partial \theta_{bi} \partial \theta_{bj}} \right) &= \text{Re}(U_{i,j} \boldsymbol{\varepsilon} G + U_i \boldsymbol{\varepsilon} V_j + U_j \boldsymbol{\varepsilon} V_i + F \boldsymbol{\varepsilon} V_{i,j})
\end{aligned}$$

and

$$\begin{aligned}
U_{i,j} &= \frac{\partial U_i}{\partial \theta_{bj}} = (D^H H_{i,j} B R^{-1} B^H D) + (D^H H_i C_j R^{-1} B^H D) \\
&\quad + (D^H H_i B L_j B^H D) + (D^H H_i B R^{-1} C_j^H D) \\
&\quad + (D^H H_j C_i R^{-1} B^H D) + (D^H \Pi_{\Pi_B A}^\perp C_{i,j} R^{-1} B^H D) \\
&\quad + (D^H \Pi_{\Pi_B A}^\perp C_i L_j B^H D) + (D^H \Pi_{\Pi_B A}^\perp C_i R^{-1} C_j^H D)
\end{aligned}$$

$$\begin{aligned}
& + (D^H H_j B R^{-1} C_i^H D) + (D^H \Pi_{\Pi_B A}^\perp C_j R^{-1} C_i^H D) \\
& + (D^H \Pi_{\Pi_B A}^\perp B L_j C_i^H D) + (D^H \Pi_{\Pi_B A}^\perp B R^{-1} C_{i,j}^H D) \\
& + (D^H H_j B L_i B^H D) + (D^H \Pi_{\Pi_B A}^\perp C_j L_i B^H D) \\
& + (D^H \Pi_{\Pi_B A}^\perp B L_{i,j} B^H D) + (D^H \Pi_{\Pi_B A}^\perp B L_i C_j^H D) \\
V_{i,j} & = \frac{\partial V_i}{\partial \theta_{bj}} = (P A^H C_{i,j} R^{-1} B^H A P)^T + (P A^H C_i L_j B^H A P)^T \\
& + (P A^H C_i R^{-1} C_j^H A P)^T + (P A^H C_j L_i B^H A P)^T \\
& + (P A^H B L_{i,j} B^H A P)^T + (P A^H B L_i C_j^H A P)^T \\
& + (P A^H C_j R^{-1} C_i^H A P)^T + (P A^H B L_j C_i^H A P)^T \\
& + (P A^H B R^{-1} C_{i,j}^H A P)^T \\
C_{i,j} & = \frac{\partial B}{\partial \theta_{bi} \partial \theta_{bj}} = 0 \\
H_{i,j} & = \frac{\partial}{\partial \theta_{bi} \partial \theta_{bj}} \Pi_{\Pi_B A}^\perp = -O_{i,j} A (A^H \Pi_B A)^{-1} A^H \Pi_B - O_i A P_j A^H \Pi_B \\
& - O_i A (A^H \Pi_B A)^{-1} A^H O_j - O_j A (A^H \Pi_B A)^{-1} A^H O_i \\
& - \Pi_B A P_j A^H O_i - \Pi_B A (A^H \Pi_B A)^{-1} A^H O_{i,j} \\
& - O_j A P_i A^H \Pi_B - \Pi_B A P_{i,j} A^H \Pi_B - \Pi_B A P_i A^H O_j \\
P_{i,j} & = \frac{\partial}{\partial \theta_{bj}} P_i = -P_j (A^H O_i A) (A^H \Pi_B A)^{-1} \\
& - (A^H \Pi_B A)^{-1} (A^H O_{i,j} A) (A^H \Pi_B A)^{-1} - (A^H \Pi_B A)^{-1} (A^H O_i A) P_j \\
O_{i,j} & = \frac{\partial}{\partial \theta_{bj}} O_i = C_{i,j} (B^H B)^{-1} B^H + C_i Q_j B^H + C_i (B^H B)^{-1} C_j^H + C_j Q_i B^H \\
& + B Q_{i,j} B^H + B Q_i C_j^H + C_j (B^H B)^{-1} C_i^H + B Q_j C_i^H + B (B^H B)^{-1} C_{i,j}^H \\
Q_{i,j} & = \frac{\partial}{\partial \theta_{bj}} Q_i = -Q_j (C_i^H B + B^H C_i) (B^H B)^{-1} - (B^H B)^{-1} (C_i^H B + B^H C_i) Q_j \\
& - (B^H B)^{-1} (C_i^H C_j + C_j^H C_i) (B^H B)^{-1} \\
L_{i,j} & = \frac{\partial R^{-1}}{\partial \theta_{bi} \partial \theta_{bj}} = -\frac{\partial R^{-1}}{\partial \theta_{bj}} G_i R^{-1} - R^{-1} \frac{\partial G_i}{\partial \theta_{bj}} R^{-1} - R^{-1} G_i \frac{\partial R^{-1}}{\partial \theta_{bj}} \\
& = -L_j G_i R^{-1} - R^{-1} G_{i,j} R^{-1} - R^{-1} G_i L_j
\end{aligned}$$

$$G_{i,j} = \frac{\partial G_i}{\partial \theta_j} = C_{i,j}^H A P A^H B + C_i^H A P A^H C_j + C_j^H A P A^H C_i + B^H A P A^H C_{i,j} \\ + \sigma_n^2 C_{i,j}^H B + \sigma_n^2 C_i^H C_j + \sigma_n^2 C_j^H C_i + \sigma_n^2 B^H C_{i,j}$$

*Proof.* The proof of this lemma is omitted because it is tedious but straightforward, utilizing matrix differentiation rules.

The gradient in (7) and Hessian in (9) contain the unknown parameter  $\theta$ , which we are ultimately interested to find, and therefore, cannot be directly used for Newton-Raphson iteration for finding  $\theta_b$ . In reality, however, information on the interval  $\Theta_s$ , where  $\theta$  must be in, is usually available to us, for example from a surveillance radar, and therefore, it is appropriate to use the averaged CRB

$$\overline{CRB}(\Theta_s, \theta_b) = \int_{\Theta_s} CRB(\theta, \theta_b) d\theta \approx \sum_{\theta_i \in \Theta_s} CRB(\theta_i, \theta_b) \cdot \Delta\theta.$$

This, in turns, gives the following averaged gradient and Hessian, after interchanging the integration and differentiation:

$$\nabla[\overline{trCRB}(\Theta_s, \theta_b)] \approx \sum_{\theta_i \in \Theta_s} \nabla[trCRB(\theta_i, \theta_b)] \cdot \Delta\theta,$$

$$H[\overline{trCRB}(\Theta_s, \theta_b)] \approx \sum_{\theta_i \in \Theta_s} H[trCRB(\theta_i, \theta_b)] \cdot \Delta\theta.$$

Therefore, the optimal  $\hat{\theta}_b$  is found by

$$\hat{\theta}_{i+1} = \hat{\theta}_i - \mu [H\{tr(\overline{CRB}(\Theta_s, \hat{\theta}_i))\}]^{-1} \nabla[\overline{trCRB}(\Theta_s, \hat{\theta}_i)]$$

where  $\hat{\theta}_i = [\hat{\theta}_{b1}, \Lambda, \hat{\theta}_{bk}]_i^T$ .

## 4. SIMULATION RESULTS

For a low-angle ship-borne radar, the target reflection usually returns back to the radar via two pathes – the direct path and the sea-surface reflection path [13, 14]. Although the AOA of the direct path is well-defined (for a continuous tracking radar), the AOA of the reflection path cannot be accurately predictable due to the diffuse refelction from rough sea surface. We shall consider a tracking radar with three beams, and obtain the optimal beam steering angles for this multipath senario.

First, we compare the CRBs of the proposed BS processing, with the DFT-based BS processing and the ES processing. Fig. 2 is the case when  $M = 16$ ,  $K = 3$ ,  $L = 6$  and  $SNR = 10dB$ . It is assumed that  $\theta_1 = 1^\circ$  and  $\theta_2 \in \Theta_s$  where  $\Theta_s = [-1.3^\circ, -1.1^\circ]$ . In this case, the computed optimal steering angles are  $\theta_{b1} = 2.7266^\circ$ ,  $\theta_{b2} = -0.6966^\circ$  and  $\theta_{b3} = -2.3956^\circ$ , while the DFT-based steering angles are  $\theta_{b1} = -7.1808^\circ$ ,  $\theta_{b2} = 0^\circ$  and  $\theta_{b3} = 7.1808^\circ$ . The CRB of the proposed BS

processing is not only lower than that of the DFT-based BS processing, but also close to that of the ES processing throughout the interval  $\Theta_s$ .

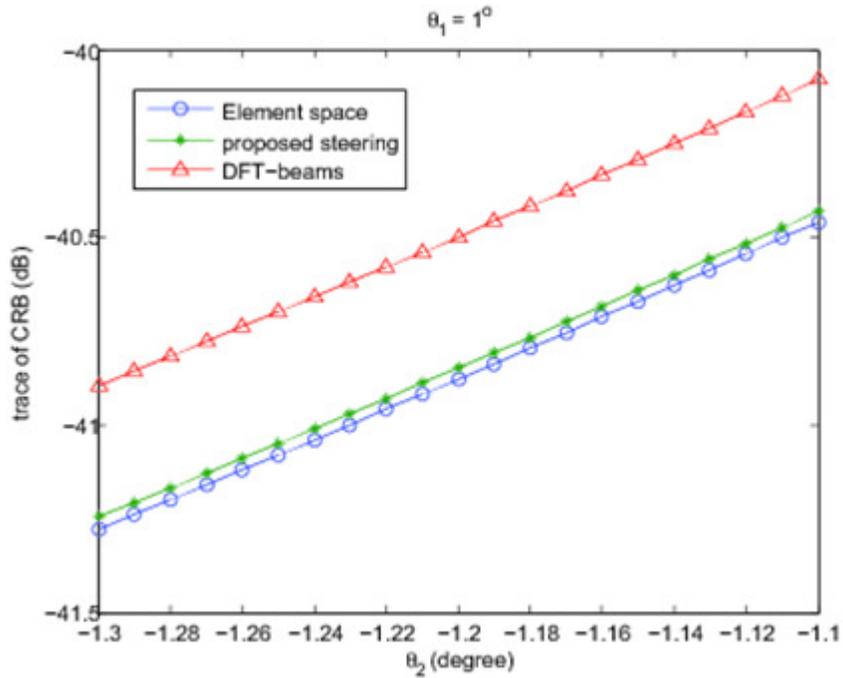


Figure 2: Comparison of the CRB with the proposed steering to that with the DFT-beam

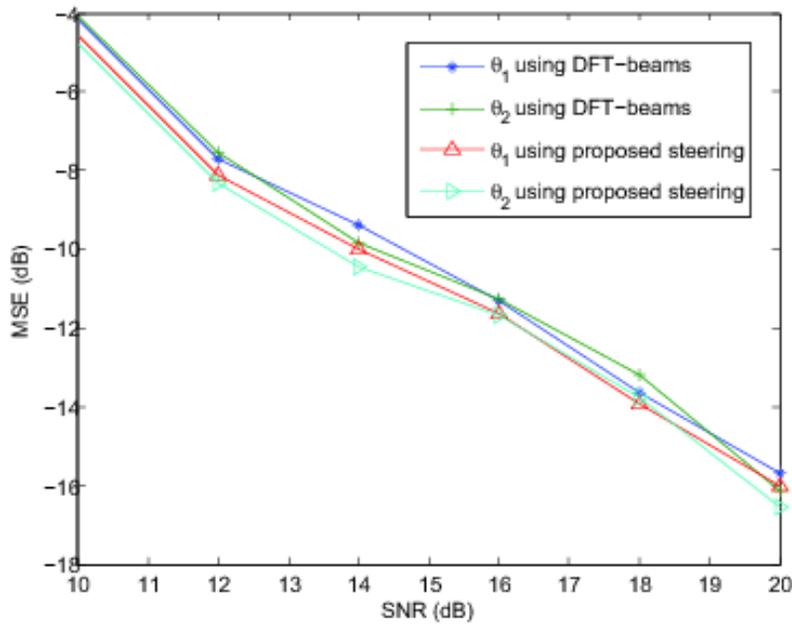


Figure 3: Comparison of the MSE with the proposed steering to that with the DFT beam

Second, we use the stochastic maximum likelihood (SMLE) method [8]

$$\begin{aligned}
\theta^{i+1} &= \theta^i - \mu_i H^{-1}(\theta^i) V'(\theta^i) \\
V'(\theta^i) &= 2 \operatorname{Re} \left( \operatorname{Diag} \left[ g(\theta^i)^H \hat{R} P_{a_s}^\perp d(\theta^i) \right] \right) \\
H(\theta^i) &= 2 \hat{\sigma}^2 \operatorname{Re} \{ d(\theta^i)^H P_{a_s}^\perp d(\theta^i) \cdot g^H(\theta^i) \hat{R} g(\theta^i) \} \\
g(\theta^i) &= a_s \left[ \left( a_s^H \hat{R} a_s \right)^{-1} - \hat{\sigma}^{-2}(\theta^i) \left( a_s^H a_s \right)^{-1} \right] \\
\hat{\sigma}^2(\theta^i) &= \frac{1}{M-N} \operatorname{Tr} \{ P_{a_s}^\perp \hat{R} \}
\end{aligned}$$

Here  $a_s = UB^H a(\theta^i)$  is evaluated for two different  $B$ s; the proposed one and the DFT-based one. Mean squared errors (MSEs) of the two estimates for 100 trials are compared in Fig. 3.

## 5. CONCLUSIONS

We have used the non-unitary beamspace CRB formula [3] to derive the gradient and Hessian of the trace of the CRB. Then applying these gradient and Hessian to a three-beam array antenna, we have found a set of three optimum beamsteering angles, for the case of two impinging plane waves, where one AOA can be rather accurately predictable, while the other has a large uncertainty. The MSE with the SMLE AOA estimation with the resulting optimum three-beam array is shown to be lower than the MSE with orthogonal DFT-beam AOA estimation.

## ACKNOWLEDGEMENTS

This work was supported in part by Hanwha Thales in 2014-2015.

## REFERENCES

- [1] Sören Anderson. On optimal dimension reduction for sensor array signal processing. *Signal Processing*, 30(2):245–256, 1993.
- [2] Russell D Brown, Richard A Schneible, Michael C Wicks, Hong Wang, and Yuhong Zhang. Stap for clutter suppression with sum and difference beams. *IEEE Trans. Aerosp. Electron. Syst.*, 36(2):634–646, 2000.
- [3] Sanghyuck Choi, Joohwan Chun, Inchan Paek, and Jonghun Jang. A stochastic crb for non-unitary beam-space transformations and its application to optimal steering angle design. *IEEE Signal Process. Lett.*, 22(11):2014–2018, 2015.
- [4] Jonny Eriksson and Mats Viberg. Adaptive data reduction for signals observed in spatially colored noise. *Signal processing*, 80(9):1823–1831, 2000.
- [5] Philippe Forster and Georges Vezzosi. Application of spheroidal sequences to array processing. In *Acoustics, Speech and Signal Processing, IEEE International Conference on ICASSP'87.*, volume 12, pages 2268–2271. IEEE, 1987.
- [6] Aboulnasr Hassanien, Sherif Abd Elkader, Alex B Gershman, and Kon Max Wong. Convex optimization based beam-space preprocessing with improved robustness against out-of-sector sources. *IEEE Trans. Signal Process.*, 54(5):1587–1595, 2006.

- [7] Minghui Li and Yilong Lu. Dimension reduction for array processing with robust interference cancellation. *IEEE Trans. Aerosp. Electron. Syst.*, 42(1):103–112, 2006.
- [8] Björn Ottersten, Mats Viberg, Petre Stoica, and Arye Nehorai. Exact and large sample maximum likelihood techniques for parameter estimation and detection in array processing. Springer, 1993.
- [9] Petre Stoica, Erik G Larsson, and Alex B Gershman. The stochastic CRB for array processing: A textbook derivation. *IEEE Signal Process. Lett.*, 8(5):148–150, 2001.
- [10] Barry Van Veen and Richard A Roberts. Partially adaptive beamformer design via output power minimization. *IEEE Trans. Acoust., Speech, Signal Process.*, 35(11):1524–1532, 1987.
- [11] Barry Van Veen and Bruce Williams. Structured covariance matrices and dimensionality reduction in array processing. In *Spectrum Estimation and Modeling, 1988.*, Fourth Annual ASSP Workshop on, pages 168–171. IEEE, 1988.
- [12] Eunjung Yang, Ravi Adve, and Joohwan Chun. Hybrid direct data domain sigma-delta space-time adaptive processing algorithm in non-homogeneous clutter. *IET Radar, Sonar and Navigation*, 4(4):611–625, 2010.
- [13] Michael D Zoltowski and Ta-Sung Lee. Maximum likelihood based sensor array signal processing in the beamspace domain for low angle radar tracking. *IEEE Trans. Signal Process.*, 39(3):656–671, 1991.
- [14] Dongmin Park, Eunjung Yang, Soyeon Ahn, Joohwan Chun, Adaptive beamforming for low-angle target tracking under multipath interference, *IEEE Trans on Aerosp. Electron Syst.* 50(4), pp. 2564-2577, 2014.

# A SWITCHED-ANTENNA NADIR-LOOKING INTERFEROMETRIC SAR ALTIMETER FOR TERRAIN-AIDED NAVIGATION

Inchan Paek<sup>1</sup>, Jonghun Jang<sup>2</sup>, Joohwan Chun<sup>3</sup> and Jinbae Suh<sup>3</sup>

<sup>1</sup>PGM Image Sensor Centre, Hanwha Thales, Kyunggi-do, Korea  
ic.paek@hanwha.com

<sup>2</sup>Agency for Defense Development, Daejeon, Korea

<sup>3</sup>Department of Electrical Engineering, KAIST, Daejeon, Korea  
chun@kaist.ac.kr

## ABSTRACT

*Conventional terrain-aided navigation (TAN) technique uses an altimeter to locate the position of an aerial vehicle. However, a major problem with a radar altimeter is that its beam (or pulse) footprint on the ground could be large, and therefore the nadir altitude cannot be estimated accurately. To overcome this difficulty, one may use the nadir-looking synthetic aperture radar (SAR) technique to reduce the along-track beam width, while the cross-track ambiguity is resolved with the interferometry technique. However, the cross-track resolution is still far from satisfactory, because of the limited aperture size of antennas. Therefore, the usual three-antenna array cannot resolve multiple terrain points in a same range bin, effectively. In this paper, we propose a technique that can increase the cross-track resolution using a large number of antennas, but in a switched fashion, not raising hardware cost.*

## KEYWORDS

*TAN, SAR, Altimeter, Switched Array*

## 1. INTRODUCTION

Terrain aided navigation (TAN) technique uses an altimeter to find the position and attitude of an aerial vehicle. However, a major problem with a radar altimeter is that its beam (or pulse) footprint on the ground could be large. For example, in Fig. 1, if the beam width is large, the leading edge of the received echo signal would not be from the nadir point  $N$ .

To overcome this difficulty, one may use the nadir-looking synthetic aperture radar (SAR) technique to reduce the along-track beam width [1], [2], while the cross-track ambiguity is resolved with the interferometry technique [3], [4], [8]. However, the cross-track resolution is still far from satisfactory, because of the limitation in the number of antennas. Therefore, the usual three-antenna array [7] cannot resolve multiple terrain points in a same range bin, effectively. We may remark that the side-looking SAR [6], which gives high resolution, both along and cross tracks, is usually not acceptable for a low-altitude TAN application. In this paper,

we propose a technique that can increase the cross-track resolution using a large number of antennas, but in a switched fashion, not raising hardware cost.

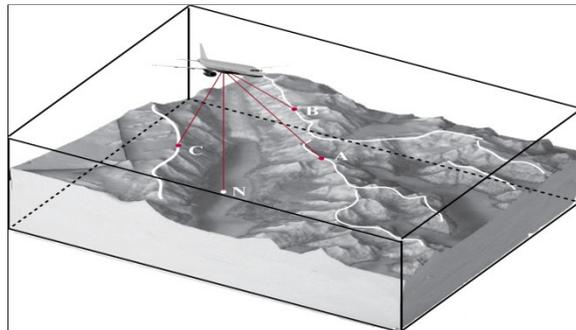


Figure 1: Multipoint-terrain-aided navigation

Figure 2 shows the flight geometry. We assume that there is a separate single transmit antenna, and  $N$  receive antennas. Also assume that there are  $L$  range bins and  $M$  coherent pulses. Then, we can find the range  $r$ , along-track angle  $\theta^a$  and the cross-track angle  $\theta^c$  to a given terrain point, respectively by measuring the echo return time, Doppler and angle of arrival (AOA).

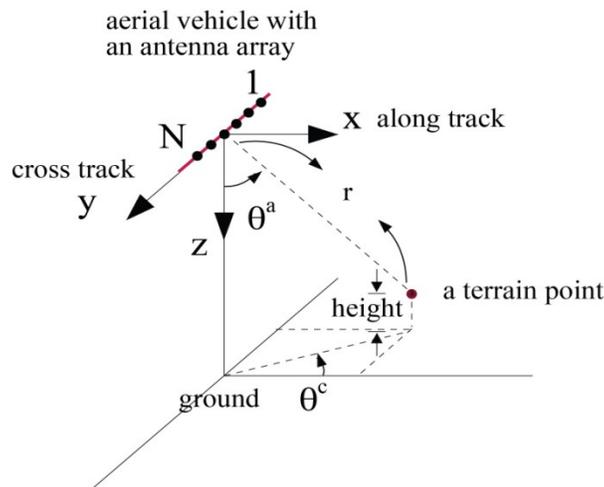


Figure 2: Flight geometry

Fig. 3 shows a 3D data cube, which represents a point target response after the range deramping operation.

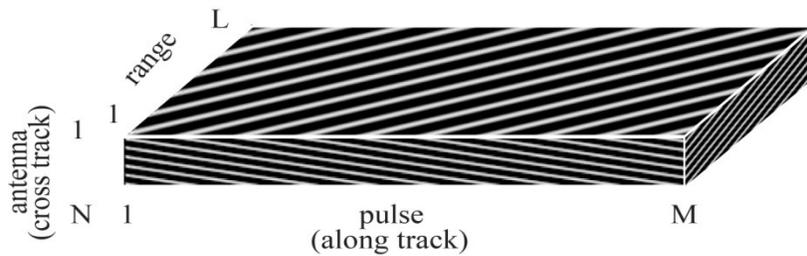


Figure 3: Deramped raw-data for a point target

The point target position in the 3D can be found by a 3D fast Fourier transform (DFT) of the data cube. The 3D data cube for actual terrain which has an infinite number of point targets will be the superposition of individual point target response, and the reconstruction of the terrain profile in Fig. 1 can be carried out by a 3D FFT of the data cube. From the reconstructed terrain profile, we can estimate the position and attitude of the aerial vehicle.

## 2. THE SWITCHED ARRAY ALTIMETER

To obtain an accurate terrain profile, the number of antennas  $N$  must be large, resulting in a large number of RF chains. However, it turns out that a reasonably accurate terrain profile can be found by selecting a small number of  $K = N$  antennas, randomly at each pulse. Fig. 4 shows the raw data when only 12% (Fig 4b) and 6% (Fig 4b) of antennas are selected.

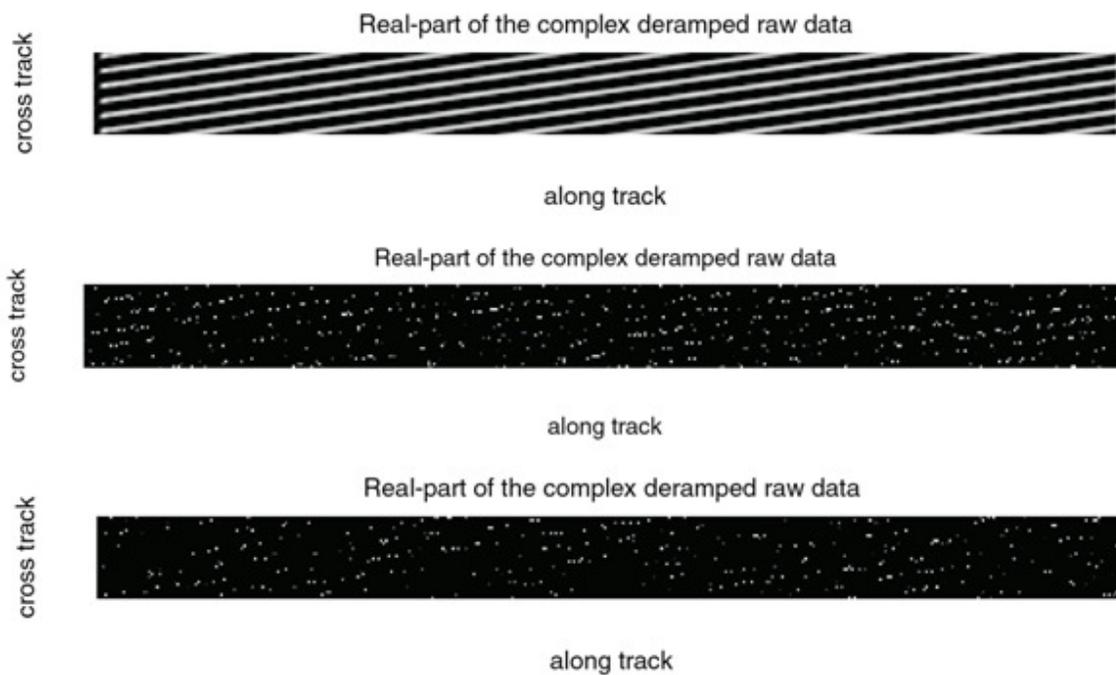
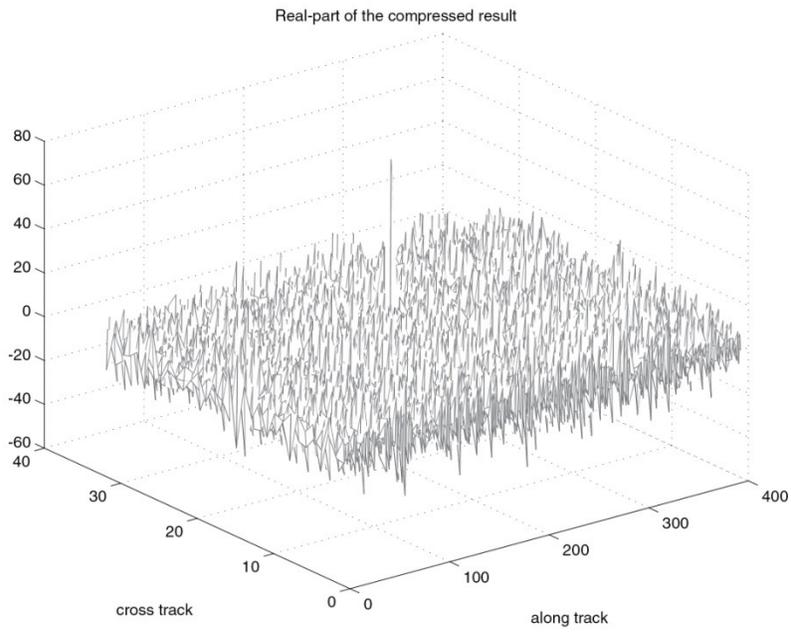
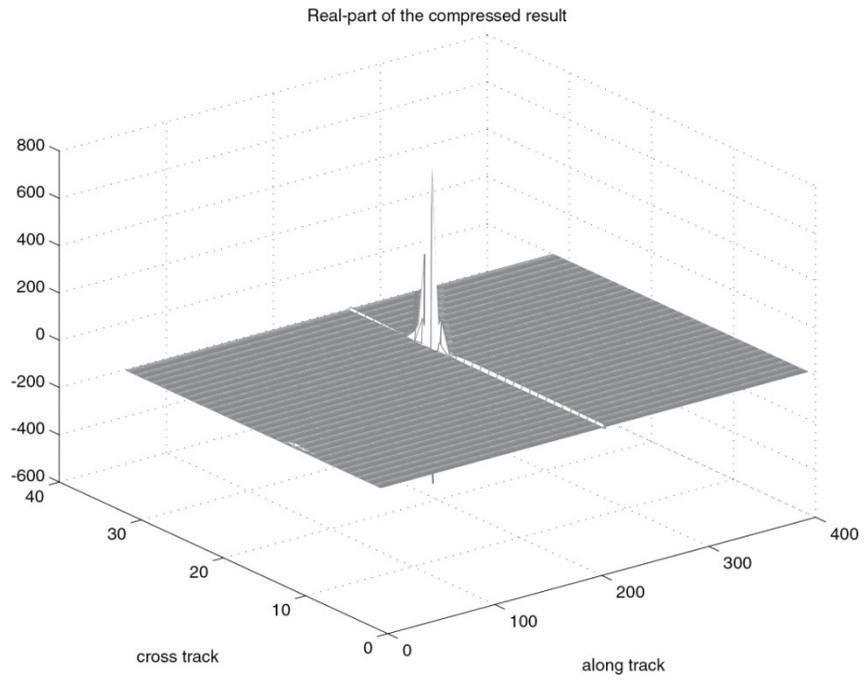


Figure 4. Raw data when (4a)  $K = N = 32$ , (4b)  $K = 3, N = 32$ , (4c)  $K = 2, N = 32$

To find the cross-track angle and range to the target, we need to compute the 2D FFTs of the raw data in Fig 4. The results are shown in Fig. 5. Note that the raw data obtained with switching produce reasonably accurate angle and range estimates.



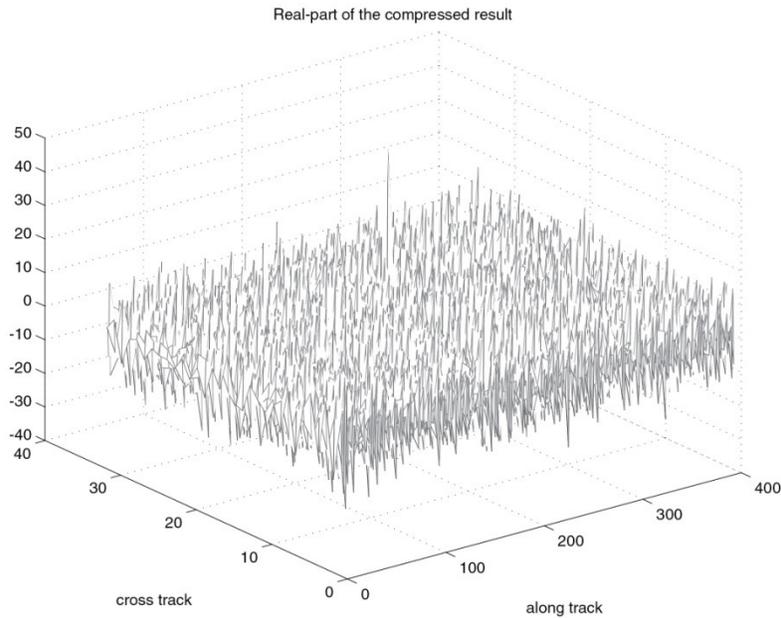


Figure 5. Recovered point target from the raw data obtained by switching. The three plots correspond to the three raw data in Figure 4.

### 3. TRACKING ALGORITHM

We use the usual 15-state Kalman filter [5], [9] to find the attitude and position of the vehicle. In particular, the time-update is carried out using the inertial measurement unit (IMU) data, and the measurement-update is done by comparing the reconstructed terrain profile with the on-board DEM (See Fig. 6). The estimated terrain profile will be particularly accurate at the mountain ridges, which are marked with white curves in Fig 1. Referring to the on-board DEM, we select multiple points (A, B, C in Fig 1) on the ridges, and determine the angles  $\theta^a$  s and  $\theta^c$  s. Then we determine the ranges  $r$  s to these points. These measurements are then compared with the predicted measurements  $\hat{f}$  to get innovation  $e$ .

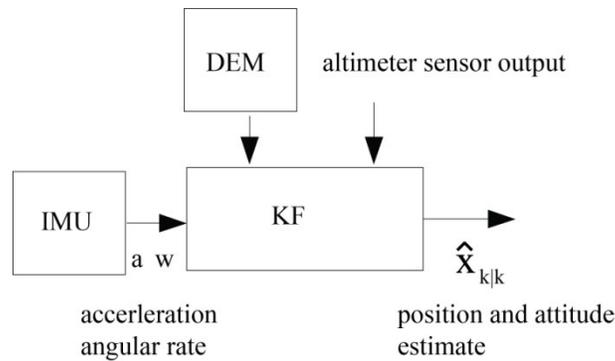


Figure 6. Block diagram of the INS

Fig 7 shows the detailed computational steps of the proposed inertial navigation system (INS). Our preliminary simulation results show that the convergence of the state estimate is quite faster with the proposed multi-point TAN than with the usual single point TAN.

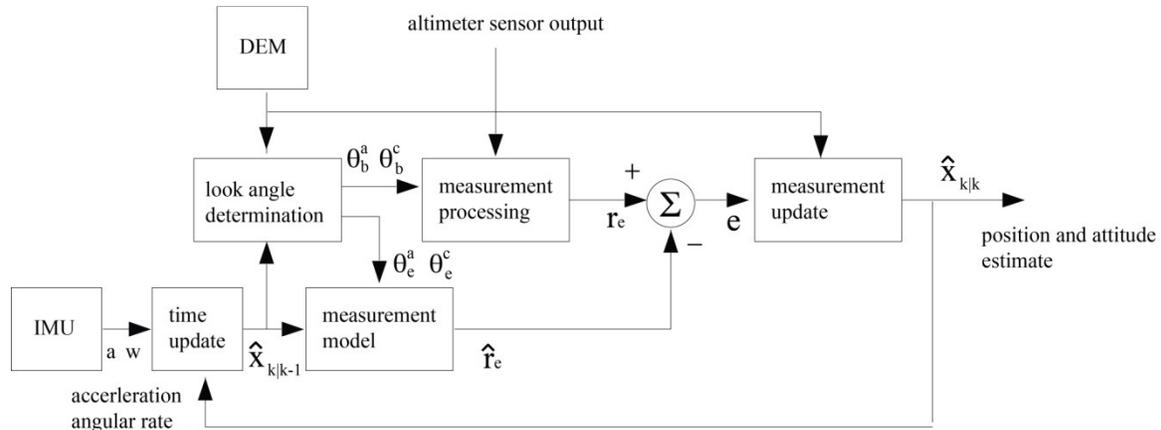


Figure 7. The proposed INS

## 4. CONCLUSIONS

We have proposed a switched antenna interferometric SAR TAN algorithm that can overcome the drawback of conventional interferometric SAR TAN. The proposed technique uses a small number of RF channels which are switched to a large number of antennas. Therefore, our proposed technique will give a high cross-track resolution comparable to the along-track SAR-based resolution, not raising hardware cost much.

## ACKNOWLEDGEMENTS

This work was supported in part by Hanwha Thales in 2014-2015.

## REFERENCES

- [1] R. Raney, "The Delay/Doppler radar altimeter," IEEE Trans. on Geosci. and Remote Sensing, Vol 36, No 5, Sep. pp. 1578-1588, 1998
- [2] R. Raney and J. Jensen, "An airborne CryoSat prototype: The D2P radar altimeter," Proceedings of the International Geoscience and Remote Sensing Symposium, Toronto, pp. 1765-1767, 2002
- [3] J. Jensen, "Angle measurement with a phase monopulse radar altimeter," IEEE Trans. On Antennas Propagat., vol 47, No 4, pp. 715-724, 1999.
- [4] J. Hager, "Interferometric synthetic aperture radar altimeter," Assignee: Honeywell Inc, US Patent 6,025,800, Feb 2000
- [5] S. Carreno, P. Wilson, P. Ridao, and Y. Petillot, "A survey on terrain based navigation for auvs," MTS/IEEE OCEANS, Seattle, 2010.

- [6] J. Curlander and R. McDonough, Synthetic Aperture Radar, Systems and Signal Processing, John Wiley, New York, 1991.
- [7] H. Jenkins, Small-Aperture Radio Direction Finding, Artech House, Boston, 1991.
- [8] J. Chun, S. Choi, I. Paek, D. Park and K. Yoo, "End-to-end design consideration of a radar altimeter for terrain-aided navigation," Proc. SPIE 8891, SAR Image Analysis, Modeling, and Techniques XIII, 889108 Oct. 2013.
- [9] D. Titterton and J. Weston, Strapdown inertial navigation technology, second eds, IEE, 2004.

*INTENTIONAL BLANK*

# OPTIMIZATION IN ENGINE DESIGN VIA FORMAL CONCEPT ANALYSIS USING NEGATIVE ATTRIBUTES

Rodríguez-Jiménez, J. M, Cordero, P, Enciso, M and Mora, A

Universidad de Málaga, Andalucía Tech,  
Boulevard Louis Pasteur SN, Málaga, Spain  
jmrodriguez@ctima.uma.es

## **ABSTRACT**

*There is an exhaustive study around the area of engine design that covers different methods that try to reduce costs of production and to optimize the performance of these engines. Mathematical methods based in statistics, self-organized maps and neural networks reach the best results in these designs but there exists the problem that configuration of these methods is not an easy work due the high number of parameters that have to be measured.*

*In this work we extend an algorithm for computing implications between attributes with positive and negative values for obtaining the mixed concepts lattice and also we propose a theoretical method based in these results for engine simulators adjusting specific and different elements for obtaining optimal engine configurations.*

## **KEYWORDS**

*Knowledge Discovery, Formal Concept Analysis, Negative Attributes, Attribute Implications, Engine Design*

## **1. INTRODUCTION**

An engine, or motor, is a machine designed to convert one form of energy into mechanical energy. There are several kinds of engines that are used in different situations. The most popular are heat engines, including internal combustion engines and external combustion engines, that burn fuel to create heat, which then creates a force. Some of these engines create electricity that is used in electric motors that convert electrical energy into mechanical motion. Engines are not only artificial machines because in nature we can find some examples in biological systems with the form of molecular motors, like myosin in muscles, that use chemical energy to create forces and eventually motion in animals and plants.

The performance of an engine measure different kind of properties like engine speed, torque, power, efficiency or sound levels. Focusing in have the best result in one of them could do that other properties have a poor result so engineers have to take into account all of these properties in general.

The cost of engine design is high so engineers have to use simulators to configure models. In the design of engines, simulators are an excellent tool to save cost of production because in them engineers can adjust different parameters and observe the global result. Complex engines could have thousands of parameters that have to be adjusted for obtaining optimal results in different ways. The problem is: what parameter they have to adjust? How these parameters have to be modified?

In general usage, design of experiments (DOE) or experimental design is the design of any information-gathering exercises where variation is present. Engineers could manage this process under full control, where statistics are usually used, or not, using artificial intelligence methods. Formal planned experimentation is often used in evaluating physical objects, structures, components and materials. Design of experiments is thus a discipline that has very broad application across all the natural and social sciences and, of course, engineering.

Datasets are necessary to collect all the information in experiments results. There are several kinds of datasets but in engine design all data is measured in numerical values. These datasets can be reduced, without lose of information, to Formal Concept Analysis compatible datasets where a binary relation determines if an object has a property or attribute. In this area we can use different methods and properties that have been developed and can not be used in the original dataset.

The mining of negative attributes from datasets has been studied in the last decade to obtain additional and useful information. There exists an exhaustive study around the notion of negative association rules between sets of attributes. However, in Formal Concept Analysis, the needed theory for the management of negative attributes is in an incipient stage. We proposed in a previous work an algorithm, based on the NextClosure algorithm [1], that allows obtaining mixed implications. The proposed algorithm returns a feasible and complete mixed implicational system by performing a reduced number of requests to the formal context.

Knowledge discovering is nowadays a well established discipline focussed on the development of tools and techniques to reveal useful information hidden in data sets. Its main goal is to detect patterns to improve decision making and is approached using pattern recognition, clustering, association and classification. Part of these patterns is expressed as implications (or association rules) which allow us to address information using a formal notation and to manage them syntactically by using logic.

In Formal Concept Analysis we have 3 related items such that we can manage them to obtain knowledge: Implications, Concepts and Context.

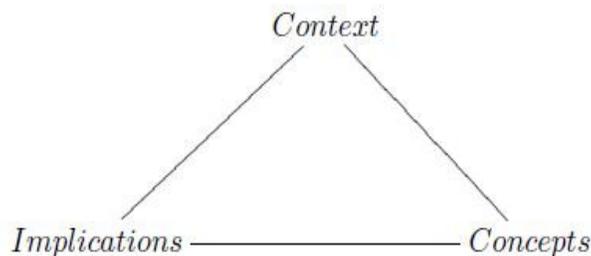


Figure 1. Items in Formal Concept Analysis

Contexts are data sets with a binary relation  $\mathbf{I}$ , between a set of objects  $\mathbf{G}$  and their attributes  $\mathbf{M}$ . Implications are formulas in the form  $A \rightarrow B$  where  $A$  and  $B$  are subsets of a certain set (universe) of attributes  $\mathbf{M}$ . Both subsets of attributes are considered to be conjunctive cubes, i.e.  $A = a_1 \wedge \dots \wedge a_n$  and  $B = b_1 \wedge \dots \wedge b_m$ .

Concepts are closed sets that have a pair formed by a subset of objects and a subset of attributes that are related with two mappings that are called derivation operators. We can calculate the closure of a subset of attributes with respect to a set of implications and establish the concept associated.

Formal Concept Analysis have a extended set of applications in real life that allows to manage information to extract knowledge from a context, a set of implications or a set of concepts: Social Networks, Marketing, Recommendation Systems,...

We are going to focus in the existence of logics developed to specify and manage sets of implications. The pioneer of these logics was the one introduced by W.W. Armstrong [2], which was proven to be sound and complete.

In this work we use implications in the area of formal concept analysis [3] and assume the common interpretation in this environment: given a formal context  $\mathbf{K}$  over a set of attributes  $\Omega$ , the implication  $A \rightarrow B$  asserts that any object which have all the  $A$  attributes, also has all the  $B$  attributes. Although we focus on knowledge discovery techniques to extract implications in formal concept analysis, this problem is similar to the extraction of functional dependencies or association rules from an arbitrary data set.

One of the former researchers who points out the importance of this problem in datasets was H. Mannila [4] and it was also studied by other researches from the database areas like S. Navathe [5]. In this work we will address the mining of implications with negation. The extended implications allow us to relate items which conflict with each other. While classical implications express that *{`cyclist with short and sharp accelerations are great sprinters"}*, implications with negations allow us to express that *{`cyclist with short and sharp accelerations are not great climbers"}*.

Since implication formulas are built using a binary connective which relates two conjunctive clauses, negation is only considered at the attribute level, i.e. the negation  $\bar{A}$  is considered the conjunctive clause of its negated attributes:  $\bar{a}_1 \wedge \bar{K} \wedge \bar{a}_n$ . We want to increase the expressiveness to exploit all the properties that Formal Concept Analysis had, but not result in much higher costs.

Notice that extended implications cannot be considered as the negation of a classical implication. We do not want to express that a certain implication does not holds but the evidence of the absence of a certain attribute because this is a solved question.

Applications of mixed attributes are usual in data mining but not in Formal Concept Analysis. There are some usual approaches, for example detecting errors [6], which provide benefits to knowledge discovery.

## 2. PRELIMINARIES

In this section, the basic notions related with Formal Concept Analysis (FCA) [7] and attribute implications are briefly presented. See [3] for a more detailed explanation.

A *formal context* is a triple  $\mathbf{K}=\langle G, M, I \rangle$  where  $G$  and  $M$  are finite non-empty sets and  $I \subseteq G \times M$  is a binary relation. The elements in  $G$  are named objects, the elements in  $M$  attributes and  $\langle g, m \rangle \in I$  means that the object  $g$  has the attribute  $m$ . From this triple, two mappings  $\uparrow: 2^G \rightarrow 2^M$  and  $\downarrow: 2^M \rightarrow 2^G$ , named derivation operators, are defined as follows: for any  $A \subseteq G$  and  $B \subseteq M$ ,

$$A^\uparrow = \{m \in M \mid \text{for each } g \in A: \langle g, m \rangle \in I\}$$

$$B^\downarrow = \{g \in G \mid \text{for each } m \in B: \langle g, m \rangle \in I\}$$

$A^\uparrow$  is the subset of all attributes shared by all the objects in  $A$  and  $B^\downarrow$  is the subset of all objects that have the attributes in  $B$ . The pair  $(\uparrow, \downarrow)$  constitutes a Galois connection between  $2^G$  and  $2^M$  and, therefore, both compositions are closure operators.

A pair of subsets  $\langle A, B \rangle$  with  $A \subseteq G$  and  $B \subseteq M$  such  $A^\uparrow = B$  and  $B^\downarrow = A$  is named a *formal concept*.  $A$  is named the *extent* and  $B$  the *intent* of the concept. These extents and intents coincide with closed sets wrt the closure operators because  $A^{\uparrow\downarrow} = A$  and  $B^{\downarrow\uparrow} = B$ . Thus, the set of all the formal concepts have a lattice structure, named *concept lattice*, with the relation

$$\langle A_1, B_1 \rangle \leq \langle A_2, B_2 \rangle \text{ if and only if } A_1 \subseteq A_2 \text{ (or equivalently, } B_2 \subseteq B_1)$$

The concept lattice can be characterized in terms of attribute implications. An *attribute implication* is an expression  $A \rightarrow B$  where  $A, B \subseteq M$  and it holds in a formal context if  $A^\downarrow \subseteq B^\downarrow$ . That is, any object that has all the attributes in  $A$  has also all the attributes in  $B$ . It is well known that the sets of attribute implications that are satisfied by a context satisfy the Armstrong's Axioms:

[Ref] Reflexivity: If  $B \subseteq A$  then  $\vdash A \rightarrow B$ .

[Augm] Augmentation:  $A \rightarrow B \vdash A \cup C \rightarrow B \cup C$ .

[Trans] Transitivity:  $A \rightarrow B, B \rightarrow C \vdash A \rightarrow C$ .

A set of implications  $B$  is an *implicational system* for  $K$  if: (1) any implication from  $B$  holds in  $K$  and (2) any implication that  $K$  satisfies follows (can be inferred) by using Armstrong's Axioms from  $B$ .

## 3. OUR APPROACH FOR EXTENDING FORMAL CONCEPT ANALYSIS WITH NEGATIVES ATTRIBUTES

From now on, the set of all the attributes is denoted by  $M$  and its elements by the letter  $m$  possibly with sub indexes. The elements in  $M \cup \bar{M}$  are going to be denoted by the first letters in the alphabet:  $a, b, c \dots$ . So, the symbols  $a, b, c \dots$  could represent positive or negative attributes. Capital letters  $A, B, C \dots$  denote subsets of  $M \cup \bar{M}$ . If  $A \subseteq M \cup \bar{M}$ , then  $\bar{A}$  denotes the set of the opposite of attributes in  $A$ . That is,  $\bar{A} = \{\bar{a} \mid a \in A\}$  where  $\bar{a} = a$ . Moreover, for  $A \subseteq M \cup \bar{M}$ , the following sets are defined:

$$\begin{aligned} \text{Pos}(A) &= \{m \in M \mid m \in A\} \\ \text{Neg}(A) &= \{m \in M \mid \bar{m} \in A\} \\ \text{Tot}(A) &= \text{Pos}(A) \cup \text{Neg}(A) \end{aligned}$$

and, therefore,  $\text{Pos}(A), \text{Neg}(A), \text{Tot}(A) \subseteq M$ .

The traditional derivation operators defined in Formal Concept Analysis are modified in [8] to consider the new framework.

**Definition 1:** *Mixed Concept-forming Operators.* Let  $K = \langle G, M, I \rangle$  be a formal context. We define the operators  $\uparrow: 2^G \rightarrow 2^{M \cup \bar{M}}$  and  $\downarrow: 2^{M \cup \bar{M}} \rightarrow 2^G$  as follows: for  $A \subseteq G$  and  $B \subseteq M \cup \bar{M}$ ,

$$A^\uparrow = \{m \in M \mid \langle g, m \rangle \in I \text{ for all } g \in A\} \cup \{\bar{m} \in \bar{M} \mid \langle g, m \rangle \notin I \text{ for all } g \in A\}$$

$$B^\downarrow = \{g \in G \mid \langle g, m \rangle \in I \text{ for all } m \in B\} \cap \{g \in G \mid \langle g, m \rangle \notin I \text{ for all } \bar{m} \in B\}$$

We prove in [9] that the adaptation of the derivation operators form a Galois connection.

**Theorem 1:** Let  $K = \langle G, M, I \rangle$  be a formal context. The pair of derivation operators  $(\uparrow, \downarrow)$  introduced in Definition 1 is a Galois Connection.

**Example 1:** Chemical analysis of oil in motors allows detecting problems inside them.

Table 1. Presence of chemical residuum in analysis of oil.

	Fe	Cu	Pb	Al	Si
$o_1$	0	1	1	1	0
$o_2$	0	1	0	1	0
$o_3$	1	0	0	1	0
$o_4$	0	0	0	0	1

$$\{o_1, o_2\}^\uparrow = \{Cu, Al\} \neq \{\bar{Fe}, Cu, Al, \bar{Si}\} = \{o_1, o_2\}^\uparrow$$

Also, we adapt the definitions of formal concept and implication to allow the use of negative attributes inside them.

**Definition 2:** Let  $K = \langle G, M, I \rangle$  be a formal context. A *mixed formal concept* in  $K$  is a pair of subsets  $\langle A, B \rangle$  with  $A \subseteq G$  and  $B \subseteq M \cup \bar{M}$  such  $A^\uparrow = B$  and  $B^\downarrow = A$ .

**Definition 3:** Let  $K = \langle G, M, I \rangle$  be a formal context and let  $A, B \subseteq M \cup \bar{M}$ , the context  $K$  satisfies a *mixed attribute implication*  $A \rightarrow B$ , if  $A^\downarrow \subseteq B^\downarrow$ .

There are special cases of mixed attributes subsets that have important properties for the proposed algorithm and reduce the complexity of it. Missaoui [10] works with intents that have empty support, i.e. have any  $m \in M$  such that  $m\bar{m}$  in the intent, so these subsets of attributes doesn't appear in the real life. We want to discard them because they are not useful for our purposes.

**Definition 4:** Let  $K = \langle G, M, I \rangle$  be a formal context and a set  $A \subseteq M \cup \bar{M}$  is named consistent set if  $\text{Pos}(A) \cap \text{Neg}(A) = \emptyset$ . The set of all consistent sets is denoted **Ctts**.

**Definition 5:** Let  $K = \langle G, M, I \rangle$  be a formal context and a set  $A \subseteq M \cup \overline{M}$  is said to be full consistent set if  $A \in \text{Ctts}$  and  $\text{Tot}(A) = M$ .

We are going to extend the logic that propose Armstrong's Axioms with a sound and complete system proposed by Missaoui with a global framework and not for particular problems.

Therefore, the following axioms are added to the Armstrong's axioms: for all  $a, b \in M \cup \overline{M}$  and  $A \subseteq M \cup \overline{M}$ ,

[Cont] Contradiction:  $\vdash a\overline{a} \rightarrow M\overline{M}$ .

[Rft] Reflection:  $Aa \rightarrow b \vdash A\overline{b} \rightarrow \overline{a}$ .

The algorithm for calculating the implicational system that we propose in [8] uses the axiomatic system and, considering the full consistent sets, the dataset have to be checked  $3^{|M|} - 2^{|M|}$  times in the worst case.

---

**Function Closed( $A, \mathfrak{B}$ ):** boolean

---

**Data:**  $A \in \text{Ctts}$ , and  $\mathfrak{B}$  being a set of mixed implications.

**Result:** 'true' if  $A$  is closed wrt  $\mathfrak{B}$  or 'false' otherwise.

```

1  begin
2  |   foreach  $B \rightarrow C \in \mathfrak{B}$  do
3  |   |   if  $B \subseteq A$  and  $C \not\subseteq A$  then
4  |   |   |   exit and return false
5  |   |   if  $B \setminus A = \{a\}$ ,  $A \cap \overline{C} \neq \emptyset$ , and  $\overline{a} \notin A$  then
6  |   |   |   exit and return false
7  |   return true
8  end
```

---

**Algorithm 2: Mixed Implications Mining**

---

**Data:**  $K = \langle G, M, I \rangle$

**Result:**  $\Sigma$  set of implications

```

1  begin
2  |    $\Sigma := \emptyset$ ;
3  |    $Y := \emptyset$ ;
4  |   while  $Y < M$  do
5  |   |   foreach  $X \subseteq Y$  do
6  |   |   |    $A := (Y \setminus X) \cup \overline{X}$ ;
7  |   |   |   if Closed( $A, \Sigma$ ) then
8  |   |   |   |    $C := A^{\uparrow\downarrow}$ ;
9  |   |   |   |   if  $A \neq C$  then
10 |   |   |   |   |    $\Sigma := \Sigma \cup \{A \rightarrow C \setminus A\}$ 
11 |   |    $Y := \text{Next}(Y)$  // i.e. successor of  $Y$  in the lexic order
12 |   return  $\Sigma$ 
13 end
```

---

#### 4. ALGORITHM FOR MIXED CONCEPTS AND MIXED ATTRIBUTE IMPLICATIONS

If we want to calculate both, mixed implicational system and set of mixed concepts, we can extend algorithm 2 [8]. We know that we only have to modify a few lines because it works with closures.

To adapt the algorithm, we have to pay attention to the condition *While*  $Y < M$  because it has to be changed to *While*  $Y \neq M$  for detecting the concepts, so the time of calculating arise. Because we only have to add full consistent sets, that are the minimal concepts that have the intents of the objects, we don't need to change this condition, and only add these intents.

If we choose the first option we have to do  $2^{|M|}$  operations that always be bigger or equal than  $|G|$ . So, the modifications are in lines 12-13, that take the closed sets that we ignore in Algorithm 1, and lines 15-16 that add to the sets of intents of concepts the full consistent sets.

---

##### Algorithm 3: Mixed Implications and Concepts Mining

---

```

Data:  $\mathbb{K} = \langle G, M, I \rangle$ 
Result:  $\Sigma$  set of implications,  $\mathcal{B}$  set of intents
1  begin
2  |  $\Sigma := \emptyset;$ 
3  |  $\mathcal{B} := \emptyset;$ 
4  |  $Y := \emptyset;$ 
5  | while  $Y < M$  do
6  | |   foreach  $X \subseteq Y$  do
7  | | |    $A := (Y \setminus X) \cup \overline{X};$ 
8  | | |   if Closed( $A, \Sigma$ ) then
9  | | | |    $C := A^{\downarrow\uparrow};$ 
10 | | | |   if  $A \neq C$  then
11 | | | | |    $\Sigma := \Sigma \cup \{A \rightarrow C \setminus A\}$ 
12 | | | | |   else
13 | | | | |    $\mathcal{B} := \mathcal{B} \cup C$ 
14 | | |    $Y := \text{Next}(Y)$  // i.e. successor of  $Y$  in the lexic order
15 |   foreach  $g \in G$  do
16 | |    $\mathcal{B} := \mathcal{B} \cup g^{\uparrow}$ 
17 |   return  $\Sigma, \mathcal{B}$ 
18 end

```

---

With this algorithm we extract all possible implications and concepts from the context. In the worst case, we have to check  $3^{|M|} - 2^{|M|} + |G|$  times the dataset that not depend of different amount of objects. The dataset could be mixed-clarified and the number of possible objects has a maximum in  $2^{|M|}$ .

#### 5. OPTIMIZING ENGINE DESIGN

In engine simulators, each component has specific measures that determine how these components work. Width of a simple pipe could increase or decrease the amount of fuel that the engine receive, so all this measures are important in the development of each engine. Each measure could be considered as an attribute of the engine and also, the relation among them determine the global efficiency. All these measures are collected in a dataset that show how the engine works but it is difficult to decide specifically which of them we can change and how we have to change it because we don't know how these changes affect engine performance. Taguchi [11] refers that development engineers have to apply parameter design methods to make the basic functions approach the ideal functions under real conditions. These design activities should be

conducted by research and development departments before the product is finally created. Taguchi's work is based in statistical methods that actually are replaced with neural networks processes or Multi-objective Covariance Matrix Adaptation Evolutionary Strategy methods [12, 13]

The method that we propose tries to focus on these specific parts of the engine that could increase the performance and change these values for obtaining the best result. Researchers could observe that is similar to a self-organized map [14], with the difference that use properties of mixed concept lattices and we only have to change a reduced number of attributes, reducing the cost of the procedure.

Our starting point is different experimental configurations of the engine that experts consider as initial results. Let  $G$  be the set of different collected configurations (objects) and  $M$  the set of elements that could be measured adding a value called *objective function* (attributes). Value of the objective function depends on different values of attributes, for example number of revolutions, but could not be associated with a particular equation. However, in the next example, for demonstration purposes, we have to consider a polynomial function.

**Example 2:** Let consider domain of possible values  $[0, 10]$  for all attributes and the objective function is  $a^2+(b-5)+(c-5)-(d-5)-e^2$ , so objective domain is  $[-115,115]$ . The attributes with higher weights are  $a$  and  $e$ , so changing these values we can modify the objective function in different ways.

Let's consider that experts determine that the optimal value for the objective is 100 and we want to know what are the preferred size for different parts of the engine for obtaining the nearest value to the proposed optimal.

Table 2: Dataset with engine configurations

T	Part a	Part b	Part c	Part d	Part e	Objective
$c_1$	3	1	7	2	3	1
$c_2$	4	3	2	3	4	-3
$c_3$	6	5	4	4	2	32
$c_4$	6	0	3	6	5	3
$c_5$	2	7	7	3	1	9

Let  $T$  be a dataset with size  $|G|*|M|$  that considers configurations in rows and measures of each element of the engine in columns. Elements in  $T$ ,  $(t_{x,y}$  with  $x \in G$ ,  $y \in M$ ) are numbers that represents each different value for the attributes but not necessary represents a formal context.

We have to point to the objective function and determine the selected value for this function that experts consider as optimal. Different values for objective function could be near from the optimal value but we need the optimal or, in cases where it is not possible, the nearest one.

A specific number  $k$ , with  $1 < k < |G|$ , have to be selected that determine the size of the control fixed group. This group  $G_k$  is compounded by  $k$  configurations which have the nearest objective values with respect to the optimal value. For each attribute  $m \in M$ , we can fix an interval  $In_m = [l_m, u_m]$  being  $l_m = \inf(t_{x,m})$  and  $u_m = \sup(t_{x,m})$  with  $x \in G_k$ .

Now we can adapt  $T$  to a compatible dataset  $T^1$  in Formal Concept Analysis in this way. Let define the values of the adapted dataset as  $t_{x,y}^1=1$  if  $t_{x,y} \in \text{In}_y=[l_y, u_y]$ ; 0 otherwise

This adaptation represents all the values that are in the fixed intervals with a binary relation. It is obvious that, for all  $g \in G_k$ ,  $m \in M$ ,  $t_{g,m}^1=1$  so we have to focus in the absence of attributes in the sense of Formal Concept Analysis. The absence of attributes means that these values are out of the fixed intervals (fixed by  $G_k$ ) and points to possible candidates for replacing their values.

**Example 3:** Fixing  $k=2$  and optimal value for objective 100 we have that  $G_k=\{c_3, c_5\}$  and  $\text{In}_a=[2,6]$ ,  $\text{In}_b=[5,7]$ ,  $\text{In}_c=[4,7]$ ,  $\text{In}_d=[3,4]$ ,  $\text{In}_e=[1,2]$

Table 3: Adapted formal context to  $T$

$T^1$	Part a	Part b	Part c	Part d	Part e	Objective
$c_1$	1	0	1	0	0	1
$c_2$	1	0	0	1	0	-3
$c_3$	1	1	1	1	1	32
$c_4$	1	0	0	0	0	3
$c_5$	1	1	1	1	1	9

The problem that we have to solve is what  $t_{x,y} \in T$  we have to change and what values could have  $t_{x,y}$ .

Using Algorithm 3 with the context  $\langle G-G_k, M, T^1 \rangle$ , we can obtain the mixed concepts lattice. The top element is the mixed concept  $\langle G-G_k, (G-G_k)^\uparrow \rangle$ . Since the lattice has an order, the upper concept that has the less number of negative attributes (not zero) determines the selected set of attributes to change. We are going to call this subset of attributes  $N$ .

For avoiding local fix points for the objective function, we have to consider 4 different values for each attribute that we are going to change in  $T$ : original value, average between minimum possible value for attribute  $y$  and  $l_y$ , average between maximum possible value for attribute  $y$  and  $u_y$ , and average for the interval  $\text{In}_y$ , for each  $y \in N$ , i.e. for each configuration, we have to consider  $4^{|N|}$  combinations.

If changing attributes in  $N$  we do not obtain better objective values, we have to follow in order the mixed concept lattice and get the next mixed concept for obtaining the new  $N$ . If proposed changes do not offer a better objective value for all the objects that are not in  $G_k$ , we have to finish our search.

This method have to be repeated recursively while distance between objective value and optimal value in  $G_k$  are higher than distance between objective value and optimal value in  $G-G_k$  after checking all possible combinations.

**Example 4:** Obtaining mixed concept lattice for  $G-G_k$ , the top element of the lattice point to the concept with attributes  $\overline{abe}$ . Parameter  $a$  is positive so it is not necessary to be changed. Focusing in negative attributes, we have to change parameters  $b, e$  in  $c_1, c_2$  and  $c_4$ , resulting Table 4.

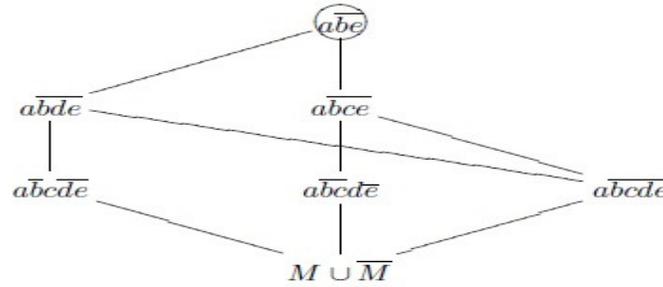


Figure 2. Mixed concept lattice for  $T^1$  restricted to  $c_1, c_2$  and  $c_4$

Table 4. Adapted formal context for T

T	Part a	Part b	Part c	Part d	Part e	Pre. Obj.	New Obj.
$c_1$	3	8.5	7	2	0.5	1	17.25
$c_2$	4	8.5	2	3	0.5	-3	18.25
$c_3$	6	5	4	4	2	32	
$c_4$	6	8.5	3	6	0.5	3	36.25
$c_5$	2	7	7	3	1	9	

We can observe that all the objectives values have changed and, in this case,  $c_1, c_2$  and  $c_4$  have an objective value higher than  $c_5$  that was in  $G_k$  in the first step, and value of  $c_4$  is higher than  $c_3$  so the distance to the optimal value decreases. Values for b increases and values for e decreases. In the next step,  $G_k=c_3,c_4$  and, recursively, we change in second step groups a, c, d; third step a, c, d; fourth step a, c, d; fifth step a, b, d, e and we find that has the same objective value that optimal selected by experts.

In 2 steps we have to check  $3*(4^2)+9*(4^3)+4^4=880$  combinations that are less than all possible combinations for each configuration, for example, if we only take into account positive integers, there are  $10^5=100000$  combinations. If we consider decimals our method is not affected but the possible combinations grow.

In this example, if we consider minimum value for y instead of average between minimum possible value for y and  $l_y$  and maximum value for y instead of average between maximum possible value for y and  $u_y$ , for each  $y \in N$ , we only have to do 112 combinations to reach the proposed optimal value, but this example have a particular solution using interval extremes, that is not the general case.

This method could be included in simulators as an artificial intelligence tool for adjusting parameters. Being an automatic tool, engineers do not need to examine and calibrate the simulator trying to detect what parameter has to change and what value is needed. Configuration for optimal value is not always reached due to parameters restrictions and the proposed original configurations, but a close value could be obtained.

As other methods, detecting local fix points is a problem that can be solved with a different set of initial configurations as a start point.

## 6. EXPERIMENTS

Due to legal problems in the use of confidential data, we can not check the method proposed in the previous section in real simulators, so we have to do simulations with different equations. These equations have been chosen for its simplicity of calculation and complexity and diversity of results according to the parameters entered.

### 6.1.Experiment 1

Table 5. Results from Experiment 1

k	G	Objective	Init. Dist.	Final Dist.	Iter.	% Reduc.
2	5	930	577.73	0.29	4.1	0.05
2	10	930	418.70	0.75	4.7	0.18
2	20	930	197.20	1.11	4.2	0.56
3	5	930	796.80	0.68	4.1	0.09
3	10	930	224.65	0.30	5.6	0.14
3	20	930	117.62	0.41	3.6	0.35
2	5	27030	14403.09	0.05	11.0	0.00
2	10	27030	14339.90	0.59	18.0	0.00
2	20	27030	10332.63	0.78	19.3	0.01
3	5	27030	15115.54	0.10	11.0	0.00
3	10	27030	13309.24	0.44	14.0	0.00
3	20	27030	13819.18	0.34	21.5	0.00
2	5	27900	19507.26	0.12	19.8	0.00
2	10	27900	15730.50	0.43	31.1	0.00
2	20	27900	14026.64	0.96	46.7	0.01
3	5	27900	16117.79	0.05	15.5	0.00
3	10	27900	17525.48	0.81	34.7	0.01
3	20	27900	11296.91	0.76	38.1	0.01

The size of  $M$  was fixed in 9 attributes which have their domains in  $[0,10]$  that allows decimal

values. The objective function for this experiment is  $\sum_{i=1}^3 t_{x,i} + (\sum_{i=4}^6 t_{x,i})^2 + (\sum_{i=7}^9 t_{x,i})^3$  for each

configuration  $x \in G$ . This function has values in  $[0,27930]$  and has 3 groups where we can interexchange values with the same result, i.e. if we change values in  $t_{x,4}$  and  $t_{x,5}$  objective function does not change.

We are going to use different values for  $k$  and  $|G|$  for the objectives 930, 27030 and 27900 that are the different configurations obtained decreasing one of the three groups to their lower values and the other groups to the upper values. 10 experiments were done with different random initial values for each fixed value for  $k$ ,  $|G|$  and objective and the average results are presented in Table 5. Initial distance is the difference in absolute distance from the best original objective value to the optimal value, final distance is the absolute distance from the best objective value to the optimal value after applying the method, iterations are the number of recursive executions of the method and reduction is the proportion between initial distance and final distance.

In this experiment we can see that, for the objective 930, there are multiple combinations of values that reach this objective and our method has difficult to locate one of them specifically due to the several fix points. For objectives 27030 and 27900, the number of possible combinations is lower, so our method can focus in particular combinations with a relevant percentage of reduction from the original distance to the optimal value. Is an interesting observation that low size of  $|G|$  has the nearest solutions to the optimal value.

## 6.2.Experiment 2

The size of  $M$  was fixed in 9 attributes which have their domains in  $[0,10]$  that allows decimal

values. The objective function for this experiment is  $\sum_{i=1}^9 (t_{x,i}-i)^2$  for each configuration  $x \in G$ .

This function has values in  $[0,485]$  and has a unique combination of values that reach the objective value 0. The difficulty to reach the optimal value in 0 is that each attribute have to get a value that is not in the extremes of its domain.

We are going to use different values of  $k$  and  $|G|$  simulating 10 experiments with different random initial values for each fixed value for  $k$  and  $|G|$ . The average results are presented in Table 6.

In this experiment, final distance is, more or less, similar in the average value of each experiment. We have a specific case where  $k=3$  and  $|G|=5$  where the solution is the nearest one to the optimal value. Since initial distance is lower when  $|G|$  increase, the percentage of reduction grows, but we can see that there is not a special relation between  $k$  and  $|G|$  in this experiment for reaching the nearest solution to the optimal value.

Table 6. Results from Experiment 2

k	G	Init. Dist.	Final Dist.	Iter.	% Reduc.
2	5	72.87	2.68	5.8	3.66
2	10	52.87	2.67	7.5	5.06
2	20	54.30	2.66	8.5	5.33
2	30	44.40	2.70	9.8	6.09
3	5	93.97	0.73	8.0	0.77
3	10	66.30	2.56	8.9	3.86
3	20	52.78	2.36	8.1	4.48
3	30	45.14	2.11	9.2	4.69
4	5	89.72	2.03	8.7	2.27
4	10	64.41	2.56	7.8	3.97
4	20	47.16	2.29	9.3	4.86
4	30	47.08	2.46	10.9	5.21

## 7. CONCLUSIONS AND FUTURE WORK

In this paper, a new algorithm for obtaining mixed concept lattices from a context is introduced. This algorithm is an important tool that let us to consider how attributes in the formal context are related among them, focusing our attention to relevant information.

The algorithm also produces an implicational system that we do not use in this work but is a complementary tool that gives us specific information about relations between attributes and was the essential idea to design the method that we propose.

This method for adjusting parameters in simulators is not a new idea, but we want to extract the knowledge hidden in datasets using formal concept analysis and, in particular, using the knowledge that negative attributes provide.

Experiments let us to check that this method could be an interesting tool for simulators with a relevant reduction in production and development costs.

In a future work, we want to develop new algorithms to reduce the time of calculating the mixed concepts lattice adapting different algorithms from Formal Concept Analysis.

Values of  $k$  and  $|G|$  have to be studied with different situations, for example changing  $|M|$ , checking what the best initial values for these parameters are.

## ACKNOWLEDGEMENTS

Supported by grant TIN2011-28084 and TIN2014-59471 of the Science and Innovation Ministry of Spain, co-funded by the European Regional Development Fund (ERDF).

We want to thanks Dr. Juan Cabrera and Dr. Juan Castillo for the approach to solving the problem discussed in this article.

## REFERENCES

- [1] Ganter, B.,(1984) "Two basic algorithms in concept analysis". Technische Hochschule, Darmstadt
- [2] Armstrong, W. W., (1974) "Dependency structures of data base relationships", Proc. IFIP Congress. Pp 580-583.
- [3] Ganter, B. & Wille, R., (1987) Formal concept analysis, Mathematical Foundations. Springer, pp. I-X, 1-284
- [4] Mannila, H, Toivonen, H. & Verkamo, A. I., (1994) "Efficient algorithms for discovering association rules", KDD Workshop, pp. 181-192.
- [5] Savasere, A., Omiecinski, E. & Navathe, S. B., (1995) "An Efficient Algorithm for Mining Association Rules in Large Databases". VLDB, pp 432-444
- [6] Revenko, A. & Kuznetsov, S. O., (2012) "Finding Errors in New Object Intents". CLA, pp 151-162
- [7] Wille, R., (1982). "Restructuring lattice theory: an approach based on hierarchies of concepts". Rival, I. (ed.), Ordered Sets, pp 445-470
- [8] Rodriguez-Jimenez, J. M., Cordero, P., Enciso, M & Mora A., (2014) "Negative attributes and implications in formal concept analysis". Procedia Computer Science, 31, pp 758-765. 2nd International Conference on Information Technology and Quantitative Management, ITQM.
- [9] Rodriguez-Jimenez, J. M., Cordero, P., Enciso, M & Mora A., (2014) "A generalized framework to consider positive and negative attributes in formal concept analysis". CLA, pp 267-278.
- [10] Missaoui, R., Nourine, L. & Renaud, Y., (2012) "Computing implications with negation from a formal context". Fundam. Inform., 115(4), pp 357-375
- [11] Taguchi, G.,(1995), "Quality engineering (Taguchi methods) for the development of electronic circuit technology". Reliability, IEEE Transactions on, 44(2), pp 225-229
- [12] Igel, C., Hansen, N. & Roth, S., (2007) "Covariance matrix adaptation for multi-objective optimization". Evol. Comput., 15(1): pp 1-28
- [13] Langouet, H., Metivier, L., Sinoquet, D. & Tran, Q., (2008) "Optimization for engine calibration"
- [14] Kohonen, T. (1982) "Self-organized formation of topologically correct feature maps". Biological Cybernetics, 43(1), pp 59-69

# COMBINED CLASSIFIERS FOR TIME SERIES SHAPELETS

Ivan S. Mitzev, Nickolas H. Younan

Mississippi State University, Mississippi State, MS 39762  
ism6@msstate.edu, younan@ece.msstate.edu

## ABSTRACT

*Time-series classification is widely used approach for classification. Recent development known as time-series shapelets, based on local patterns from the time-series, shows potential as highly predictive and accurate method for data mining. On the other hand, the slow training time remains an acute problem of this method. In recent years there was a significant improvement of training time performance, reducing the training time in several orders of magnitude. This work tries to maintain low training time- in the range from several second to several minutes for datasets from the popular UCR database, achieving accuracies up to 20% higher than the fastest known up to date method. The goal is achieved by training small 2,3-nodes decision trees and combining their decisions in pattern that uniquely identifies incoming time-series.*

## KEYWORDS

*Data mining, Time-series shapelets, Combining classifiers*

## 1. INTRODUCTION

The time-series shapelets classification method was introduced by Ye and Keogh [1] as a new type of data mining method, that uses the local features of time-series instead of their global. That makes it less sensitive to obstructive noise [1]. This method is successfully applied to a variety of application areas benefiting from its short classification time and high accuracy. Despite its advantages it has a significant disadvantage- a very slow training time. Current research mostly focuses on searching shapelets from all possible combination of time-series derived from a dataset [1, 2], keeping the training process relatively slow. A variety of proposals have been introduced to reduce the candidate shapelets [1, 2, 4, 5], but training time is still in the range of hours for some datasets. A newly introduced method [7] shrinks significantly the training time, making the training process to last from portion of a second to several seconds for investigated 45 datasets from UCR collection [9]. Although, this is the fastest up to date training method as of our knowledge, it also maintains high accuracies in compare with other state-of-arts methods [7]. In this paper we introduce a new method that reaches higher accuracies compared with the method from [7], keeping the training time in observable limits. We tested with 24 datasets from [9], the ones with number of classes higher than five. It was found that proposed method outperforms in terms of accuracy the method from [7] for most datasets. The achieved training time is kept low, varying from several seconds to several minutes, depending on a dataset. High accuracy and relatively short training time makes proposed method very competitive to present state-of-arts methods, which lack either accuracy or have huge training time.

The rest of this paper is organized as follows. In section 2 related work is presented. Section 3 describes the proposed method and gives technical details of its implementation. Section 4 discusses achieved results. Finally, section 5 summarizes the proposed method and gives ideas for further work.

## 2. RELATED WORK

Shapelet by definition is a sequence of samples that originate from one of the time-series from a dataset and maximally represent certain class. The classical method of shapelets discovery, known as *brute force algorithm* [1], employs all possible subsequences from all time-series from the train dataset and treat them as candidate shapelets. To test a candidate shapelet how well separates two classes A and B, all distances between the candidate shapelet and time-series from A and B are formed. These distances are ordered into a histogram and the histogram is consecutively split into two parts until the best information gain is achieved. The split point is named optimal split distance and distances below it considered to belong to class A, but above it to class B. If any other candidate shapelet achieves higher information gain, it is selected as shapelet. The process continue until all the candidate shapelets are processed. The method requires vast amount of calculation time. First improvements include *subsequence distance early abandon* of calculated distances and *admissive entropy pruning* based on predicted information gain [1]. These improvements reduced the total required time for training, but the reduction was not that significant [6]. Another idea based on the *infrequent shapelets*, prunes the non-frequent candidate shapelets [4]. More improvements [2] suggests using of so called *logical shapelets*, that reuse the computation and optimize the search space. Recent approach is based on synthesizing shapelets from random sequences, using *particle swarm optimization* techniques [6]. A new development in the area [7], vastly improves the training time of the shapelets by pruning candidate segments, which shows similarity in Euclidian distance space. This approach [7] is the fastest up to date as of our knowledge and in terms of accuracy is competitive with the current state-of-arts methods. We selected this method as a reference to proposed method, aiming to achieve similar or better accuracies, maintaining relatively low training time.

## 3. PROPOSED METHOD

Our previous research [6] changes the traditional way of producing shapelets by synthesizing a shapelet from randomly generated sequences using particle swarm optimization (PSO), instead of extracting the shapelets from the original time-series. It finds the shapelet for every pair of classes presented in a dataset, then combines them in a decision tree and find the decision tree that achieves highest accuracy. Producing the most accurate decision tree requires all possible combinations of trees to be tested. That is a slow process for more than four classes and adds additional processing time to traditionally slow training time. For this purpose, only datasets with less than five classes were investigated in [6]. Datasets with more than five classes are processed with the method introduced in this paper. The proposed method utilizes groups of small (up to 4 classes) decision trees, instead of building one big decision tree that contains all classes. When a time-series comes for classification every present tree produce a decision. The *decision path* taken during classification is present as string of characters. The decision paths from all present trees are combined into *decision pattern*. Every class from the datasets appears to maintain unique decision pattern. The patterns from training datasets are kept and when new time-series from test dataset comes for classification these patterns are compared with the pattern produced

from incoming time-series. The incoming time-series is associated with the class, to which its decision pattern mostly match.

### 3.1. Training

#### 3.1.1. Extracting subsets

The first step of the training process is to extract subsets of classes out of the original dataset, for which the decision trees will be defined. It is best to have uniform distribution of class indexes into subsets, as it allows non dominant class indexes into the final solution. The maximum amount of subsets  $L$  is defined as:

$$L = K! / (K - n)! n! \quad (1)$$

where,  $K$  is the number of all classes in a dataset and  $n$  is the number of classes in a subset  $n = 2, 3, 4$ .

In case the number of classes in the original dataset is relatively high (Fig.1,  $K = 37$ ), and the subsets consists of four classes for example, the final number of possible subsets combinations becomes  $66045$ , according (1). That is vast amount of subsets and training all of them will defeat the purpose of simplifying the calculations process. Instead of taking all possible combinations we can operate with just limited amount of subsets, obeying the rule of uniform distribution of class indexes as shown on Fig.1. Taking limited amount of subsets will not always fully obey the uniform rule. For example, on Fig.1 most of classes are present 3 times, but class 21, 29 and 30 are present just two times. Practically, it is enough all class indexes to be present into the subsets and the difference between the number of times classes are present to be no more than one.

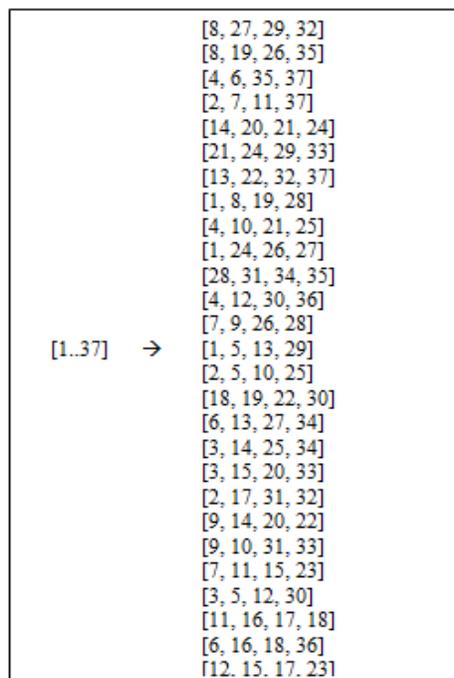


Fig. 1 Extracted subsets of 4-class combinations from a dataset with class indexes [1..37].

### 3.1.2. Training decision trees

Next step of the training process is to create decision trees for extracted subsets. Method from [6] is applied for these subsets as they consist of up to four classes. The method in [6] applies Particle Swarm Optimization (PSO) techniques, treating candidate shapelets as particles, which form a swarm. To find a shapelet between two classes A and B, swarm of  $N-2$  candidate shapelets is formed. The candidate shapelet represent a random sequence of samples and  $N$  is the length of the time-series in a dataset. Every such sequence has different length, varying from 3 (the smallest meaningful shapelet length) up to  $N$ . These potential candidates cover the whole range of possible shapelets lengths. These candidate shapelets compete to each other to find the best solution- a sequence that maximally separates two classes A and B. On every step of the process the candidate shapelet changes its values according to the best overall values in the swarm and candidates best values so far. The fitness function applies functionality similar to the criterions of the brute force algorithm. After calculating the distances between the candidate shapelet and time-series from class A and class B, it builds a histogram of distances and calculate the possible highest information gain. If currently calculated information gain is bigger than information gain assigned so far to the candidate, current values becomes particle's best values. If currently calculated information gain is bigger than information gain of the best candidate shapelet, then current candidate shapelet become the best candidate of overall swarm. Iteration stops when pre-defined number of iterations is achieved or when best information gain from iteration to iteration remains the same.

Described particle searching process is relatively fast, but high number of candidate shapelets may slow down the training process in general. For that, two improvements were introduced. First, compression of the training data is introduced, which according to [7] will not harm the training process, but improves the performance. Second, we introduce the idea that not  $N-2$  candidate shapelets are required, but only 10 will be enough to compete and find the best shapelet among them. As candidate shapelets have different lengths it is important to know which of them to remove from the competition. To find the ten most representative shapelets lengths, we extract from the training dataset only a few time-series per class and train with them. In this partial training process we use  $N-2$  shapelets candidates. The partial training is very fast as just few time-series were used, but it shows well which are the most popular shapelets lengths for certain subset. Based on these results, we select the 10 most popular shapelets lengths and run the process again.

When all pair of classes in a subset have their shapelets discovered, then all possible variations of decision trees for this subsets are checked and the one that produces the highest accuracy is selected. The accuracy of the decision trees during the training process is checked with time-series from the training dataset.

### 3.1.3. Decision patterns

An important term in proposed method is the *decision path*. The decision path is the path taken through the decision tree during decision process. When time-series comes for classification, the distance between shapelet and the time-series is calculated. If such distance is higher than the optimal split distance associated with the shapelet, the process takes the right branch of the tree, if not- the process takes the left branch. When right branch is taken, character "R" is added to the decision path, when the left branch is taken, character "L" is added to the decision path. The path

length is equal to the tree depth. An example of possible decision paths are shown on Fig.2. To form a *decision pattern* the decision paths from all decision trees are concatenated as shown on Fig. 3. For example, if the system consists of 6 decision trees and every tree has two nodes similar to that on Fig.2, then one possible variant of decision patterns is {RL,RR,R-,L-LL,RL}. During the training process, decision pattern for every time-series from the training dataset is collected. These decision patterns are kept and used during the classification process. Keeping the decision patterns into memory requires certain amount of memory to be allocated during the classification process. For example, “*Non.FatalECG.1*”[9] dataset contains 1800 train time-series and the chosen pattern length is 836 characters. If we consider that a character is encoded with 1 byte then required memory during classification process is in the vicinity of 1.5MB. That is feasible amount for the modern computers, but may cause difficulties in small embedded systems. The direction (“R/L/-”) of the decision path currently requires 1 bytes (8 bits) of encoding, but it could be optimized to two bits to reduce the required amount of memory, especially in the embedded systems.

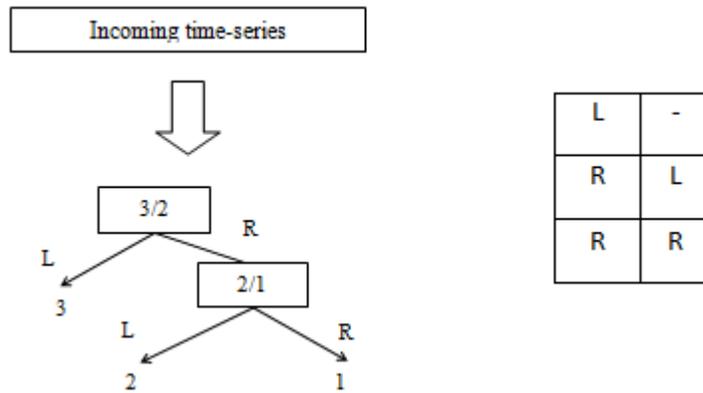


Fig.2 Possible decision paths of the illustrated decision tree obtained after classification of incoming time-series

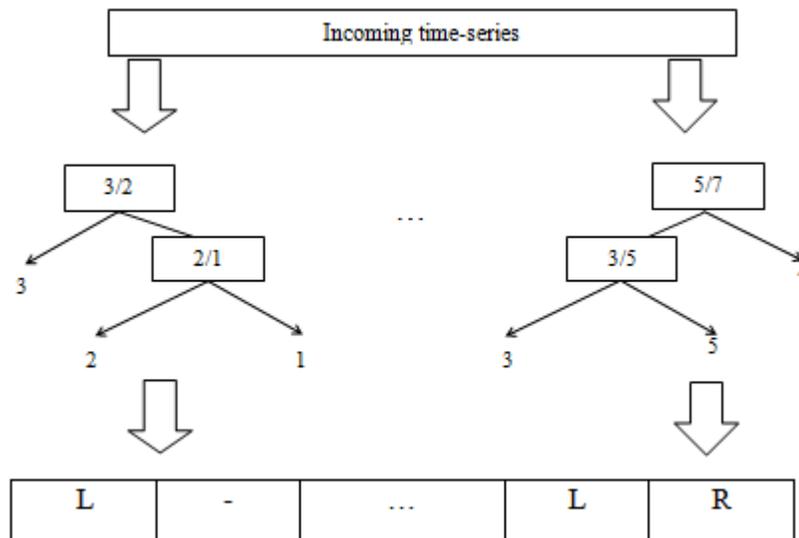


Fig.3 Decision pattern, obtained by combining the decision paths from all subsets' decision trees.

### 3.2. Classification

When time-series from test dataset comes for classification a decision pattern is created for that time-series. This pattern is compared with the patterns produced during the training process. Comparison process is very simple. The two decision patterns strings are compared character by character in place and the comparison coefficients is equal to the number of characters that coincide by place and value, divided by the length of the decision string as shown on Fig. 4. After all the comparison coefficients are collected we keep only those which are above certain threshold of  $0.98$ . The idea of this classification is that time-series from the same class will produce similar decision patterns, but decision patterns from different classes will differ a lot. The incoming time-series is identified as class to which it has closest decision pattern. In certain cases more than one class index show similar pattern to the investigates time-series. In such cases the classification process assign the incoming time-series to the class, that has majority of decision patterns closest to the incoming time-series decision pattern.

R	-	L	L	R	-	L	L	L	R
R	L	L	L	L	-	L	R	L	L

Fig. 4 Comparison between decision patterns of two time-series. Six out of ten characters coincide by place and value, therefore the comparison coefficients is  $6/10 = 0.6$ .

## 4. EXPERIMENTAL RESULTS

The project implementation uses C# and .NET Framework 4.0. Time performance measurements were produced with a *System.Diagnostics.StopWatch* .NET class. In our experiments we used a PC with the following parameters: CPU: Intel Core i7, 2.4GHz; RAM: 8 GB; 64-bit Windows 7 OS. We selected datasets from the UCR collection [9] with a number of classes higher than five (Table 1) as for datasets with fewer classes applying proposed method is meaningless. Table 1 shows parameters of the used datasets. We used method from [7] as a reference method. It produces the fastest training time as of our knowledge and accuracies that outperform most of state-of-arts method for the moment. We downloaded the Java implementation of the proposed method from [10] and ran it on the same hardware as proposed method. Reference method requires to specify threshold  $p$  and aggregation ratio  $r$ . We kept these value the same as in [7] to maintain the highest accuracy. Table 3 shows the results of both methods in terms of training time and accuracies they produce. In 18 out of 24 cases the proposed method outperforms the reference method in terms of accuracy, where the improvements vary from 2% up to 23%, where in six of these cases the improvement is above 10%. In the rest, 3 cases differ less than 1.0% and we consider that both method perform equally in this cases. Only in 3 cases the reference method outperform the proposed method in terms of accuracy, but the difference is less than 2%. Although the reference method shows better training times, the proposed method maintains an observable training time- varying from several seconds up to several minutes (~15 min. for Non.FatalECG.1) for datasets that have long time-series and higher number of time-series in a train datasets (uWave.X, uWave.Y, uWave.Z).

Proposed method uses the decision patterns from all time-series in the training datasets thus, in some cases the classification process may slow done due to high number of time-series in the train dataset (Non.FatalECG.1, Non.FatalECG.2). In some cases the length of the decision pattern

could be relatively long (Adiac- 1473, Non.FatalECG.1- 836) which may also influence the classification time. To investigate this issue we have measured the averaged classification time per time-series. Table 2 represents the results. In the most heavy case (Non.FatalECG.1) when the number time-series is 1800 and the decision pattern string length is 836 characters the classification time is above 200 milliseconds. The length of the decision pattern may vary as shown on Table 3. For datasets, such as “Beef”, which consist of 5 classes, the number of subsets is limited to 10 when constructed of 3 class indexes or to 5 when constructed of 4 class indexes. In this case to achieve better accuracy, combination of all possible trees up to 4 indexes are taken. On the other hand, datasets with more class indexes have more varieties to choose from. In the case of “50Word” dataset, which contain 50 class indexes, the total amount for combinations for two-classes decision tree is 1225. We selected 497 of them based on the principle from 3.1.1 and the total length of the decision pattern become 994 characters. Rising the number of characters in the decision pattern in all investigated cases increased the accuracy in general. Although, it appears that there is certain limit of characters above which the accuracy does not increase and even may decrease as shown on Table 4.

Table 1. Used datasets from UCR database.

Dataset	Number of Classes	Number of time-series in the train/test dataset	Time-series length
Beef	5	30 / 30	470
Haptics	5	155 / 308	1092
OsuLeaf	6	200 / 242	427
Symbols	6	25 / 995	398
synthetic.	6	300 / 300	60
Fish	7	175 / 175	463
InlineSkate	7	100 / 550	1882
Lighting7	7	70 / 73	319
MALLAT	8	55 / 2345	1024
uWave.X	8	896 / 3582	315
uWave.Y	8	896 / 3582	315
uWave.Z	8	896 / 3582	315
MedicalImages	10	381 / 760	99
Cricket X	12	390 / 390	300
Cricket Y	12	390 / 390	300
Cricket Z	12	390 / 390	300
FaceAll	14	560 / 1690	131
FacesUCR	14	200 / 2050	131
SwedishLeaf	15	500 / 625	128
WordsS.	25	267 / 638	270
Adiac	37	390 / 391	176
Non.FatalECG.1	42	1800 / 1965	750
Non.FatalECG.2	42	1800 / 1965	750
50words	50	450 / 455	270

Table 2. Averaged classification times produced by proposed method.

Dataset	Classification Time, [msec]	Dataset	Classification Time, [msec]
Beef	0.38	MedicalImages	4.27
Haptics	1.05	Cricket X	11.48
OsuLeaf	1.67	Cricket Y	9.07
Symbols	0.72	Cricket Z	9.07
synthetic.	1.81	FaceAll	10.06
Fish	3.97	FacesUCR	3.97
InlineSkate	4.14	SwedishLeaf	16.03
Lighting7	2.17	WordsS.	17.75
MALLAT	2.11	Adiac	64.56
uWave.X	3.98	Non.FatalECG.1	223.52
uWave.Y	5.87	Non.FatalECG.2	195.99
uWave.Z	3.84	50words	66.77

Table 3. Experimental results presenting accuracies and training times of the proposed and reference method.

Dataset	Comp. Rate	Proposed method			Reference method	
		Pattern Length	Train Time, [sec]	Accuracy, [%]	Train Time, [sec]	Accuracy, [%]
Beef	0.125	70	4.15	<b>52.21</b>	0.05	48.89
Haptics	0.500	20	70.66	<b>39.39</b>	1.69	34.56
OsuLeaf	0.125	150	55.84	<b>76.99</b>	0.14	53.31
Symbols	0.250	150	7.87	<b>94.20</b>	0.05	82.48
synthetic.	0.250	150	125.58	98.88	0.06	98.44
Fish	0.250	287	102.51	<b>90.85</b>	0.15	75.05
InlineSkate	0.125	245	78.57	39.57	0.56	39.88
Lighting7	0.500	245	42.52	<b>75.79</b>	0.39	65.30
MALLAT	0.125	280	42.87	<b>92.85</b>	0.10	90.77
uWave.X	0.250	117	559.22	75.32	4.37	<b>76.45</b>
uWave.Y	0.250	168	594.66	65.12	3.33	<b>66.72</b>
uWave.Z	0.125	117	508.93	66.30	1.89	<b>67.48</b>
Med.Images	0.500	240	139.47	<b>71.27</b>	0.58	67.68
Cricket X	0.250	471	267.66	<b>77.78</b>	0.61	68.63
Cricket Y	0.250	408	198.58	<b>79.14</b>	0.50	64.01
Cricket Z	0.250	414	184.38	<b>75.29</b>	0.66	68.21
FaceAll	0.500	342	167.02	<b>75.42</b>	1.25	71.63
FacesUCR	0.500	330	36.97	<b>90.56</b>	0.32	84.61
SwedishLeaf	0.500	519	342.27	<b>91.14</b>	0.34	85.60
WordsS.	0.250	600	28.48	<b>65.46</b>	0.29	60.92

Adiac	0.500	1473	514.63	<b>73.65</b>	0.27	55.67
Non.FatalECG.1	0.250	836	878.18	<b>85.01</b>	6.90	80.93
Non.FatalECG.2	0.125	836	349.32	<b>89.19</b>	4.67	86.34
50words	0.250	994	58.15	68.79	0.35	68.06

Table 4. Accuracy dependency from decision pattern length for “Lighting7”( 7 classes) dataset.

Decision Pattern Length (number of trees x number of classes in tree )	Training Time, [sec]	Accuracy, [%]
21x2	4.09	61.64
35x3	16.28	71.23
35x4	30.90	71.68
35x4 + 35x3	42.52	<b>75.79</b>
35x4 + 35x3 + 21x2	44.29	73.97

## 5. CONCLUSION AND FUTURE WORK

This paper proposes a new method for time-series shapelets classification, which demonstrate higher accuracies than produced by fastest known state-of-arts method for most of the investigated cases. As well it keeps an observable time for training, varying from several seconds to several minutes.

As future work we will focus on improving the training time even further, applying parallel processing based calculations (utilizing *.NET Parallel.ForEach*) that employ all possible processor’s cores on certain machine. This technology could be successfully applied on variety of places in proposed method where the calculations from current stage are independent from each other, namely: iteration steps of the PSO algorithm for shapelets discovery; the comparison between incoming time-series pattern and available decision patterns.

## REFERENCES

- [1] L. Ye and E. Keogh, “Time series shapelets: a new primitive for data mining,” in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.
- [2] A. Mueen, E. Keogh, and N. Young, “Logical-shapelets: an expressive primitive for time series classification,” in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011.
- [3] T. Rakthanmanon and E. Keogh, “Fast shapelets: A scalable algorithm for discovering time series shapelets,” Proceedings of the 13th SIAM International Conference on Data Mining, 2013.
- [4] He1 Q., Dong Z., Zhuang F., Shang T., Shi Z., “Fast Time Series Classification Based on Infrequent Shapelets”, 2012 11th International Conference on Machine Learning and Applications, 2012
- [5] J. Yuan, Z. Wang, H. Meng, „A discriminative Shapelets Transformation for Time Series Classification“, International Journal for Pattern Recognition and Artificial Intelligence, Vol. 28, No. 6, 2014.

- [6] I. Mitzev, N. Younan, (2015), "Time Series Shapelets: Training Time Improvement Based on Particle Swarm Optimization", 7th International Conference on Machine Learning and Computing, March 2015
- [7] J. Grabocka, M. Wistuba, L. Schmidt-Thieme, "Scalable Discovery of Time-Series Shapelets", arXiv:1503.03238 [cs.LG], March 2015
- [8] J. Lines, L. Davis, J. Hills, A. Bagnall, "A Shapelet Transform for Time Series Classification", Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012
- [9] E. Keogh, Q. Zhu, B. Hu, H. Y., X. Xi, L. Wei, and C. A. Ratanamahatana, "The UCR Time Series Classification/Clustering Homepage," [www.cs.ucr.edu/~eamonn/time\\_series\\_data](http://www.cs.ucr.edu/~eamonn/time_series_data)
- [10] J. Grabocka, M. Wistuba, L. Schmidt-Thieme, Source Code and Executables for Scalable Discovery of Time-Series Shapelets algorithm, <https://www.dropbox.com/sh/btiee2pyn6a989q/AACDfzkkpdYPmgw7pgTgUoeYa>
- [11] P.Senin, S.Malinchik, "SAX-VSM: Interpretable Time Series Classification Using SAX and Vector Space model", Data Mining (ICDM), 2013 IEEE 13th International Conference, 2013
- [12] D. Gordon, D. Hendler, L. Rokach, "Fast Randomized Model Generation for Shapelet-Based Time Series Classification", arXiv:1209.5038 [cs.LG], 2012
- [13] J. Grabocka, N. Schilling, M. Wistuba, L. Schmidt-Thieme, "Learning Time-Series Shapelets", KDD'14, August 24–27, 2014, NY, USA, 2014

## AUTHORS

**Ivan S. Mitzev** is currently PhD candidate of Electrical and Computer Engineering at Mississippi State University. He received his M.S. degree of Electrical Engineering from Mississippi State University in 2010. His research interests include software development, pattern recognition and bio-medical signal processing.



**Nicolas H. Younan** is currently the Department Head and James Worth Bagley Chair of Electrical and Computer Engineering at Mississippi State University. He received the B.S. and M.S. degrees from Mississippi State University, in 1982 and 1984, respectively, and the Ph.D. degree from Ohio University in 1988. Dr. Younan's research interests include signal processing and pattern recognition. He has been involved in the development of advanced signal processing and pattern recognition algorithms for data mining, data fusion, feature extraction and classification, and automatic target recognition/identification.



**Dr. Younan** has published over 250 papers in refereed journals and conference proceedings, and book chapters. He has served as the General Chair and Editor for the 4th IASTED International Conference on Signal and Image Processing, Co-Editor for the 3rd International Workshop on the Analysis of Multi-Temporal Remote Sensing Images, Guest Editor, Pattern Recognition Letters, and JSTARS, and Co-Chair, Workshop on Pattern Recognition for Remote sensing (2008-2010). He is a senior member of IEEE and a member of the IEEE Geoscience and Remote Sensing society, serving on two technical committees: Image Analysis and Data Fusion, and Earth Science Informatics (previously Data Archive and Distribution). He also served as the Vice Chair of the International Association on Pattern Recognition (IAPR) Technical Committee 7 on Remote Sensing (2008-2010), and Executive Committee Member of the International Conference on High Voltage Engineering and Applications(2010-2014).

# RESILIENT INTERFACE DESIGN FOR SAFETY-CRITICAL EMBEDDED AUTOMOTIVE SOFTWARE

Harald Sporer, Georg Macher, Christian Kreiner and Eugen Brenner

Institute of Technical Informatics,  
Graz University of Technology, Graz, Austria  
{sporer, georg.macher, christian.kreiner, brenner}@tugraz.at  
<http://www.iti.tugraz.at/>

## **ABSTRACT**

*The replacement of the former, purely mechanical, functionality with mechatronics-based solutions, the introduction of new propulsion technologies, and the connection of cars to their environment are just a few reasons for the continuously increasing electrical and/or electronic system (E/E system) complexity in modern passenger cars. Smart methodologies and techniques have been introduced in system development to cope with these new challenges. A topic that is often neglected is the definition of the interface between the hardware and software subsystems. However, during the development of safety-critical E/E systems, according to the automotive functional safety standard ISO 26262, an unambiguous definition of the hardware-software interface (HSI) has become vital. This paper presents a domain-specific modelling approach for mechatronic systems with an integrated hardware-software interface definition feature. The newly developed model-based domain-specific language is tailored to the needs of mechatronic system engineers and supports the system's architectural design including the interface definition, with a special focus on safety-criticality.*

## **KEYWORDS**

*Embedded Automotive Systems, Hardware-Software Interface, Model-Based Design, Domain-Specific Modelling, Functional Safety*

## **1. INTRODUCTION**

Electrical and/or electronic systems (E/E systems) in the automotive domain have grown increasingly complex over the past decades. New functionality, mainly realized through embedded E/E systems, as well as the growing connectivity (Car2X-Communication), will keep this trend alive in the upcoming years. Well-defined development processes are crucial for managing this complexity and achieving high quality products. Wide-spread standards and regulations, such as Automotive SPICE® and ISO 26262, provide guidance through the development life cycle. Some of the key aspects of these concepts are full traceability and consistency between the different development artifacts.

In the automotive industry, the E/E system architectural design models are usually created with techniques based on the *Unified Modeling Language (UML)*. Either the meta-model is extended, Jan Zizka et al. (Eds) : CCSIT, SIPP, AISC, CMCA, SEAS, CSITEC, DaKM, PDCTA, NeCoM - 2016 pp. 183–199, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.60117

or a profile is created to make it possible to use the UML-based approach in embedded automotive system design. A wide-spread example of an UML2 profile is the *Systems Modeling Language (SysML)*, which reuses many of the original UML diagram types (*State Machine Diagram, Use Case Diagram, etc.*), uses modified diagram types (*Activity Diagram, Block Definition Diagram, etc.*), and adds new ones (*Requirement Diagram, Parametric Diagram*) [1].

Even if the UML-based methodologies are valuable for projects with an emphasis on software, they are sometimes too powerful for embedded automotive system design, due to the numerous representation options. Particularly for domain experts who have no or limited knowledge of software development, the large number of elements available for modelling, turns system architectural design into an awkward task. However, it is not the intention of this work to decry the SysML approaches created so far. They are a good choice for a multitude of tasks. Instead, this paper showcases an extension of these SysML approaches, which makes the architectural design process easier, placing a special focus on the specification of the hardware-software interface for UML non-natives.

A model-based domain-specific language and domain-specific modelling (DSM) has been developed for the specific needs of embedded automotive mechatronics systems. Additionally, a software tool has been created to support the new modelling techniques. By linking development artifacts such as requirements (e.g. technical system requirements, software requirements, etc.), and verification criteria to the design model, the traceability mentioned earlier is assured.

The main goal of this work is to contribute to the improvement of the existing system architectural design methods by facilitating the specification of the hardware-software interface. The approach presented has mainly been created for the development of embedded mechatronics-based E/E systems in the automotive field. However, the techniques are also suitable for other domains. Improvements have been made by extending the system modelling approach presented in previous publication using HSI specification capabilities.

Section 2 presents an overview of related approaches, domain-specific modelling and integrated tool chains. Section 3 provides a description of the proposed hardware-software interface specification approach for the model-based system engineering. An application of the methodology described is presented in Section 4. Finally, this work is concluded in Section 5, which gives an overview of the presented work.

## **2. RELATED WORK**

In recent years, a lot of effort has been made to improve the model-based automotive E/E system design methods and techniques. Today, the advantages of a model-based approach are clear and without controversy. Meseguer [2] grants much more reliability, reusability, automatisation, and cost effectiveness to software that is developed with modelling languages. However, model transformation within or across different languages is crucial to achieve all these benefits.

Traceability and consistency between the development artifacts have always been important topics. However, these properties have become even more important due to the increasing number of electronic and electric-based functionalities. According to the international standard ISO 26262 [3], released in 2011, traceability between the relevant artifacts is mandatory for safety-critical systems. A description of the common deliverables relevant to automotive E/E

system development, and a corresponding process reference model is presented by the de facto standard Automotive SPICE [4]. Neither the functional safety standard nor the process reference model enforces a specific methodology for how the development artifacts have to be created or linked to each other. However, connecting the various work products manually is a tedious and error-prone task.

One of the early work products found in the engineering process is the system architectural design. In the field of automotive E/E system development, a wide-spread and common approach is to utilize a UML-based technique for this design, such as the UML2 profile SysML. Andrianarison and Piques [5], Boldt [6], and many other publications (e.g. [7], [8], [9]) present their SysML methodologies for system design. As stated by Broy et al. [10], the drawbacks of the UML-based design are still the low degree of formalization, and the lack of technical agreement regarding the proprietary model formats and interfaces. The numerous possibilities of how to customize the UML diagrams and how to get a language for embedded system design, are behind these drawbacks. Even if there is an agreement to utilize a common UML profile such as SysML, there are plenty of design artifact variations. This scenario does not provide an optimal base for the engineer who has to design the embedded automotive system from a mechatronics point of view. Ideally, the tool should be intuitive and it should be possible to use it easily without specific knowledge of UML.

Mernik et al. [11] describe a domain-specific language as a language that is tailored to the specific application domain. This tailoring should lead to a substantial increase in expressiveness and ease of use, compared to general-purpose languages. Even if expressiveness is increased by the utilization of SysML-based modelling techniques, the ease of use for embedded automotive mechatronics system design has not been improved.

Preschern et al. [12] claim that DSLs help to decrease system development costs by providing developers with an effective way to construct systems for a specific domain. The benefit in terms of a more effective development has to be greater than the investment needed to create or establish a DSL at a company or in a department. In addition, the authors argue that the mentioned DSL development cost will decrease significantly over the next few years, due to new tools that support language creation such as the Eclipse-based *Sirius*<sup>1</sup>.

Vujovic et al. [13] present a model-driven engineering approach to creating domain-specific modelling (DSM). Sirius is the framework used to develop a new DSM and the DSM graphical modelling workbench. The big advantage of this tool is that the workbench for the DSM is developed graphically. Therefore, knowledge about software development with Java, the graphical editor framework (GEF) or the graphical modelling framework (GMF) is not needed. Although it is obvious that an unambiguous specification of the various signals between the items of an embedded automotive system design is vital, publications on embedded automotive hardware-software interface definition are rare. This contribution aims to extend a model-based development approach for an ISO 26262 aligned hardware-software interface definition presented by the authors of [14]. More background on the origin of HSI characteristics is presented and the model-based support is shifted from a classic SysML-based methodology to a domain-specific modelling methodology for the E/E system architectural design of mechatronics-based systems. The domain-specific modelling (DSM) language definition is presented in [15].

---

<sup>1</sup> <https://eclipse.org/sirius/>

### 3. APPROACH

The main goal of this contribution is to convey the importance of the hardware-software interface for today's *Embedded Automotive Systems* and how it is supported by the approach described. Moreover, the key driving factors for establishing a well-defined interface, which is also suitable for safety-critical applications, will be shown within this section. Before describing the HSI specification approach in detail, the utilized domain-specific model-based system architectural design technique shall be introduced. This domain-specific modelling method has been developed to outline mechatronics-based system architectures in the automotive sector and therefore serves as a basis for the specification of the hardware-software interface found in our approach.

#### 3.1. Embedded Mechatronics System Domain-Specific Modelling

The key objective of domain-specific modelling is to provide a lean approach for engineers to facilitate embedded automotive mechatronics system modelling on a high abstraction level. The approach described focusses on the model-based structural description of the E/E system under development. Additionally, the signals and interfaces are an essential part of modelling.

The existing SysML-based design method (see also [14]) is extended by the newly developed *Embedded Mechatronics System Domain-Specific Modeling (EMS-DMS)* for automotive embedded system architectural design. It is not intended to replace the SysML-based solution created so far. Instead, the EMS-DMS is integrated into existing methods. Hence, the whole tool-chain, starting from the SysML-based system architectural design tool and finishing at software / hardware architectural design, can be utilized if desired. An overview of the tool integration is shown in Figure 1.

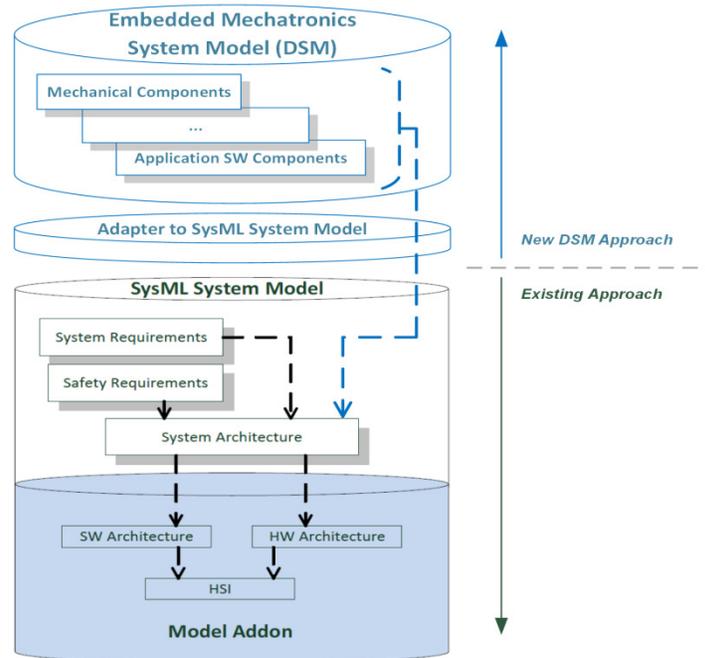


Figure 1. Tool-Chain Integration of DSM and SysML Model Approach (based on [16])

The definition of the newly developed model-based domain-specific language is shown in Figure 2. The *EMS-DSM Component* is the origin of all other classes regarding language definition. The six attributes of this class are

- *ID* - unique identifier of the particular instance in the architectural design model, set automatically.
- *Name* - name or short description of the particular instance, chosen by the design engineer.
- *Mask* - graphical representation of the particular instance, set by the engineer responsible for the design tool.
- *Requirement* - in this approach, a link to the Redmine requirements database is set by the designer.
- *Verification Criteria* - similar to Requirement, a link to the Redmine verification criteria artifact is set by the designer.
- *Specification* - link to further information about the actual component, e.g. a CAD drawing or a data sheet.

The *EMS-DSM Component* serves as the base node of the EMS-DSM definition, and declares the common attributes of the derived classes at the lower levels. Therefore, this component is not instanced for the design process. At the next language definition level, the following component classes are available:

- *Mechanical Components* - used by all mechanical, domain-specific components, e.g. the *Mechanical Pressure Regulator* class in the use-case shown in Section 4.
- *Compartment Components* - gives the opportunity to specify areas or compartments, where mechanical and hardware components are installed.
- *E/E Item Components* - an abstract component class definition, which serves as a basis for the hardware and software components at the lower levels. Additionally, the property *ASIL*, corresponding to the ISO 26262, is stated.

The majority of the non-abstract component classes are derived from the hardware component class:

- *Sensor Component* - used for all domain-specific sensor components.
- *Control Unit Component* - used for all domain-specific control unit components.
- *Actuator Component* - used for all domain-specific actuator components.
- *External Control Unit Component* - special class, to make signals from an external system available in the considered system.

All hardware components and their instances in the system design model, with the exception of the *External Control Unit Component*, are capable of containing a software design model. This means that any kind of software component instance is only allowed to be implemented in a software design model which belongs to an instance of a hardware component. This special language characteristic is defined by the *Aggregation* relationship between hardware and software components, which also implies the hardware-software interface.

The last part of the EMS-DSM definition description is related to the classes (derived from the software component):

- *Basis Software Component* - used for all low-level, hardware-dependent software components.
- *Application Software Component* - used for all functional software components.

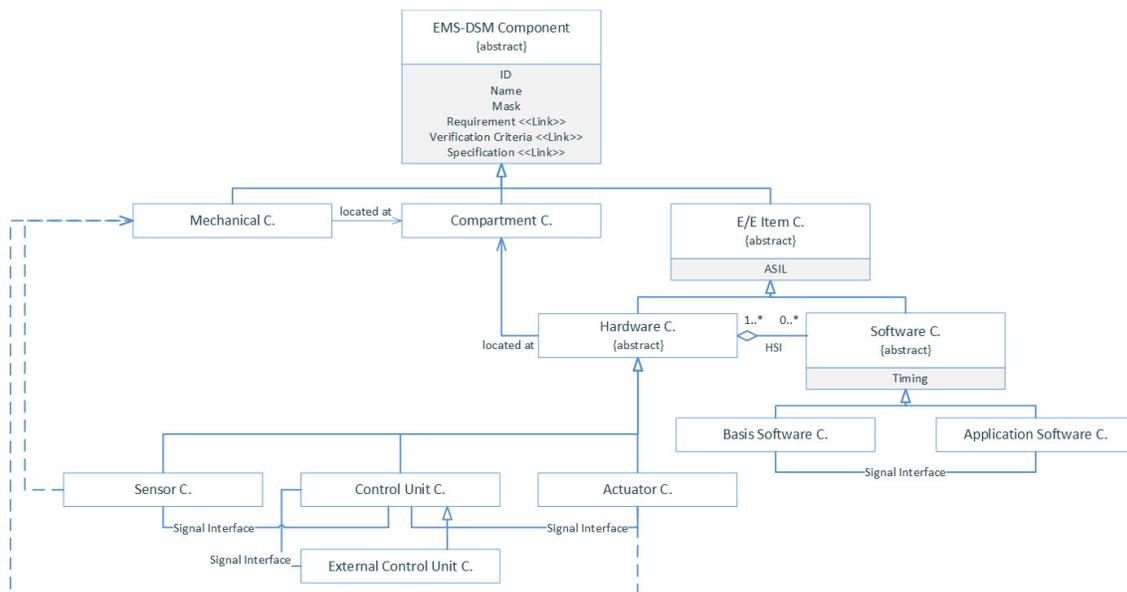


Figure 2. EMS-DSM Language Definition

As mentioned in Section 2, a more detailed description of the domain-specific modelling language can be found in [15].

### 3.2. Influence of Process Reference Model on HSI Specification

Due to their broad dissemination in the automotive sector, the two most important reference models are Automotive SPICE [4] and CMMI [17]. Both pursue similar targets: they (a) determine the process capability/maturity, and (b) aspire a continuous process improvement in the particular development team and/or company. The reference models do not exist in order to specify how processes have to be implemented. Instead, desired process outcomes (Automotive SPICE) or goals (CMMI) are defined and described in more detail by best practice characterisation (base or generic practices at Automotive SPICE, and specific or generic practices at CMMI). The *Automotive S(oftware) P(rocess) I(mprovement) and C(apability) (D)e(termination)* reference model is based on the international standard ISO 15504 and is

primarily used in Europe, as well as in some parts of Eastern Asia. The latest version, which was analysed for this approach, is 3.0 and was released in July 2015. The *C(apability) M(aturity) M(odel) I(ntegration)* reference model has been developed by the Software Engineering Institute (SEI) at Carnegie Mellon University. CMMIs currently exist for *Acquisition, Development, and Services*. As CMMI is not widespread in the European automotive sector, the remaining part of this section will focus on Automotive SPICE as the relevant process assessment and reference model. The model does not address the demand for a hardware-software interface directly, but some guidance on HSI specification can be extracted from general interface topics.

Table 1 lists the elements of the Automotive SPICE reference model that provide information about interfaces between system components. As expected, interface work products are needed for *Architectural Design* and the *Integration* topics. In addition to the *Process ID* and the *Process Name*, the corresponding *Base Practice IDs* are indicated. These give more detailed information on what the outcome should look like. In SYS.3.BP3, the definition (*identify, develop, and document*) of system element interfaces is stipulated. This equally applies to the hardware-software interface. In SYS.3.BP4, a description of the dynamic behaviour of and between the system elements is provided. The possible operating modes of the system, which determine the dynamic behaviour, have to be taken into account in the HSI definition. Base Practice SYS.4.BP3 postulates that the interfaces between system items have to be covered by the system integration test to show consistency between the real interfaces and the architectural design. With regard to the HSI, SWE.2.BP3 and SWE.2.BP4 can be interpreted in a similar way to their system level counterparts (SYS.3.BP3, SYS.3.BP4). SWE.2.BP5 claims the determination and documentation of the resource consumption objectives of all relevant software architectural design elements. To support this using the hardware-software interface definition, information on resource consumption shall be included in the description of the signals, wherever applicable. An interface definition is also demanded at process *SWE.3 - Software Detailed Design and Unit Construction*. However, in this case, the specification belongs to the signals communicated between the components on the lowest (most detailed) software level. Hence, this communication specification does not directly belong to the hardware-software interface, and will not be taken into consideration in this approach. The last process/base practice in Table 1 is SWE.5.BP3. It demands a description of the interaction between relevant software units and their dynamic behaviour. Again, this base practice can be interpreted in a similar way to its system level counterpart (SYS.4.BP3).

Table 1. HSI Accompanying Automotive SPICE Processes.

Process ID	Process Name	Base Practice ID
SYS.3	System Architectural Design	BP3, BP4
SYS.4	System Integration and Integration Test	BP3
SWE.2	Software Architectural Design	BP3, BP4, BP5
SWE.5	Software Integration and Integration Test	BP3

In the Automotive SPICE reference model, *Output Work Products* are also defined and linked to the base practices previously stated. From this contribution's point of view, the relevant work products are:

- *System Architectural Design* - the main aspects to consider regarding the HSI are *memory/capacity requirements, hardware interface requirements, security/data*

*protection characteristics, system parameter settings, system components operation modes, and the influence of the system's and system component's dynamic behaviour.*

- *Interface Requirement Specification* - the main aspects to consider regarding the HSI are *definition of critical timing dependencies or sequence ordering and physical interface definitions.*

### 3.3. Influence of Automotive Functional Safety on HSI Specification

The international standard *ISO 26262* for *Functional Safety* in the automotive electrical and/or electronic system domain was released in 2011. Since then, many best practice articles and books have been published on how to develop according to the standard. However, with the exception of the safety-critical view, the hardware-software interface has rarely been highlighted in these publications.

According to *ISO 26262*, the HSI is to be specified during the phase *Product Development at the System Level* (see Figure 3), which is described in Part 4 of the standard. As a prerequisite for specifying the hardware-software interface, a system design has to be established. While preparing the system architectural design, the technical safety and non-safety requirements are allocated to the hardware and software. Subsequent to this allocation, an initial interface description can be prepared. The HSI shall be continuously refined in the ensuing hardware and software product development phases, which are described in Parts 5 & 6 of the *ISO 26262*.

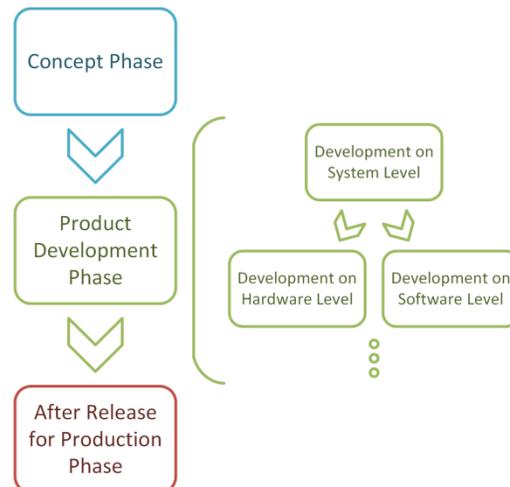


Figure 3. Development Phases According to [3]

The majority of information concerning how to specify the interface aligned to functional safety can be found in Clause 7.4.6 of Part 4 of the standard. In our approach, most of the HSI characteristics demanded by this clause, such as *operation modes of the hardware device* and *shared/exclusive use of the hardware resource*, are described in the *Detailed Hardware Specification (DHS)* documents, which are linked to the main HSI document. A detailed description of the various development artifacts and their relationships is presented in Subsection 3.4. Additionally, the informative *Annex B* of Part 4 of *ISO 26262* provides information concerning the possible content of the interface definition.

### 3.4. Incorporated Hardware-Software Interface Specification

Two main objectives have to be achieved when developing a new HSI specification approach:

1. identification, development and documentation of the essential HSI specification attributes & characteristics, and
2. support for the linking of related information to ensure full traceability.

The principle of the hardware-software interface specification approach described here is based on three origins, two of which have been described in the previous subsections:

- a. the process reference and assessment model *Automotive SPICE*,
- b. the automotive functional safety standard *ISO 26262*, and
- c. the industrial experience of authors in past automotive E/E system development projects.

It is important to note that the hardware-software interface specification does not only consist of a single spreadsheet with a description of all signals between hardware and software. Further information belonging to the HSI specification can also be found in various development artifacts. Figure 4 shows the different aspects of our HSI specification approach:

- *Hardware-Software Interface Signal List* - spreadsheet with data of all signals between hardware and software. The attributes describing each signal have been derived from sources (a) - (c), which were mentioned at the beginning of this subsection.
- *Resource Consumption Objectives* - depending on the particular project, the objectives are described in spreadsheet(s) and/or free text document(s). Regardless of the type, the documents are linked to the software components in software architectural design (see attribute *Specification <<Link>> Software Component* class in the EMS-DSM language definition in Figure 2).
- *Detailed Hardware Specification* - depending on the particular project, the objectives are described in spreadsheet(s) and/or free text document(s). Regardless of the type, the documents are linked to the hardware components in system architectural design (see attribute *Specification <<Link>> Hardware Component* class in the EMS-DSM language definition in Figure 2).
- *Model-based Architectural Design* - this item represents the central source of information. The defined domain-specific modelling language facilitates the creation of the system and the software architectural design within the same design environment and allows the linking of all other relevant development artifacts. From a HSI specification perspective, the three previous items in this list are the most important development artifacts to be linked to the architectural design models.

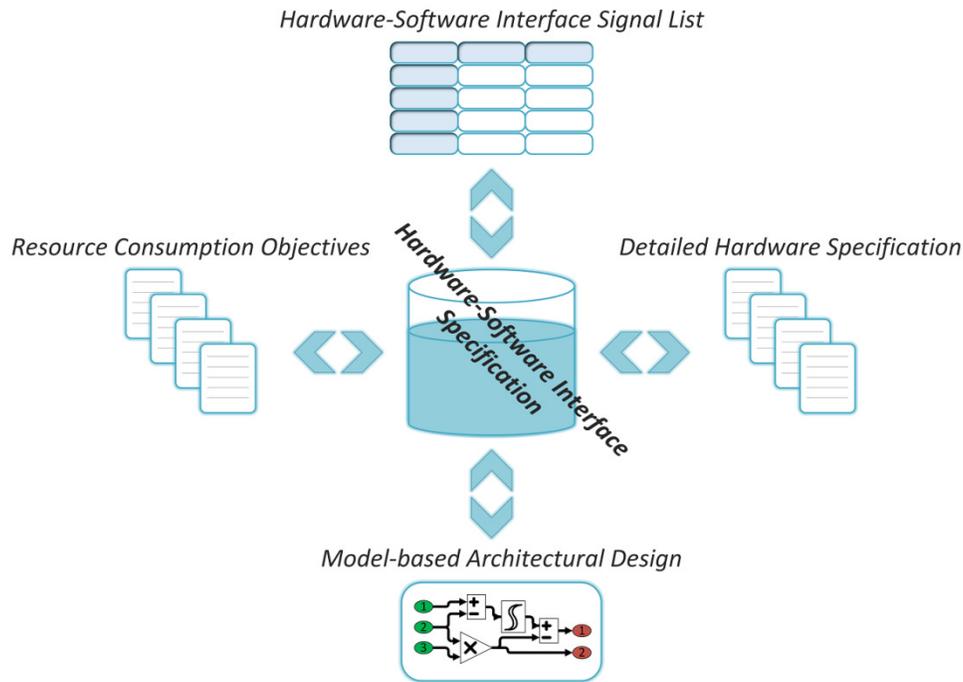


Figure 4. Distributed Hardware-Software Interface Specification

Establishing full traceability between the *Resource Consumption Objectives*, the *Detailed Hardware Specification*, and the *Model-based Architectural Design* is an easy task, accomplished by linking the related documents in the architectural design.

The integration of the *Hardware-Software Interface Signal List* data into the design model is more technically challenging. In [14] the authors described the functionality of the *HSI Definition Exporter and Importer*, which was developed to achieve a seamless transformation of the HSI representation between the SysML-based architectural design and the spreadsheet tool. The *HSI Definition Exporter* is an extension (*dynamic link library*) for the model-based development (MBD) tool, which is written in C# and allows the modelled HSI to be exported to a spreadsheet document (either in *csv* or *xls* format). The *HSI Definition Importer* is the counterpart of the *HSI Definition Exporter*, which is also implemented as a *dynamic link library* using the spreadsheet tool's API. It allows the import of all HSI information from the spreadsheet document or a selective update of the HSI model artifacts. Using both the export and import functionality leads to a round-trip engineering capability regarding the HSI signal list and the HSI signals modelled in the architectural design. In this approach, the libraries of the exporter and importer extensions are slightly adapted to the needs of the domain-specific modelling language.

To conclude the description of our approach, the HSI signal attributes and their origins are listed in Table 2.

Table 2. HSI Signal List Attributes.

Attribute	Comments	Origin
Signal Direction	Input or Output, out of the controllers view	Author's Experience
Signal Description	A short signal description or the signals name	ISO 26262-4 (Annex B)
Sensor / Actuator	Type or identifier of signals source/sink	Author's Experience
Supply Voltage	-	Author's Experience
Physical Min Value	-	ASPICE SYS.4.BP3
Physical Max Value	-	ASPICE SYS.4.BP3
Accuracy	In % of range of values	ISO 26262-4 (Annex B)
Physical Unit	E.g. V, A, ...	ISO 26262-4 (Annex B) ASPICE SYS.4.BP3
HW Interface Type	E.g. Digital In, Analog Out, CAN, ...	ISO 26262-4 (Annex B) ASPICE WP 17-08
HW Pin #	Pin number or identifier at e.g. ECU	ISO 26262-4 (Annex B)
Message ID	In case of bus communication	Author's Experience
Start Bit		
Internal Cycle Time	E.g. 10 ms	ISO 26262-4 (Section 7.4.6) ASPICE SYS.4.BP3, SWE.5.BP3, WP 17-08
External Cycle Time	Only applicable for digital signals	Author's Experience
HW Timer / Interrupt / Watchdog	Identifier of triggered e.g. interrupt	ISO 26262-4 (Section 7.4.6)
Operating Modes	Information if signal is needed special operating modes (e.g. start up, calibration, ...)	ISO 26262-4 (Annex B) ASPICE SYS.3.BP4, SWE.2.BP4, WP 04-06
HW Diagnostic Feature	E.g. short circuit detection, ...	ISO 26262-4 (Section 7.4.6)
Memory Type	E.g. RAM, EEPROM, ...	ISO 26262-4 (Annex B)
Security/Data Protection	Information on special security issues	ASPICE WP 04-06
Critical Timing Dependencies or Sequence Ordering	-	ASPICE WP 17-08
Signal Name @ SW	Identifier of signal as used in application software	Author's Experience
Initial Value	-	Author's Experience
Data Type	E.g. UInt16, Float, ...	ASPICE SYS.4.BP3, SWE.5.BP3
Scaling LSB	Scaling information in case of fixed-point arithmetic	ASPICE SYS.4.BP3, SWE.5.BP3
Scaling Offset		
Min Value @ SW	-	ASPICE SWE.5.BP3
Max Value @ SW	-	ASPICE SWE.5.BP3
Accuracy @ SW	In % of range of values	ISO 26262-4 (Annex B)
Physical Unit @ SW	E.g. km/h, Nm, ...	ASPICE SWE.5.BP3
Default Value @ SW	Default value in case of an invalid input signal	Author's Experience
Detection Time	Time until a fault is diagnosed	ISO 26262-4 (Section 7.4.6)
Reaction Time	Admissible reaction time after a fault was detected	ISO 26262-4 (Section 7.4.6)
ASIL	Automotive Safety Integrity Level classified A - D, or QM if no safety-relevance is given	ISO 26262-4 (Annex B)
Signal ID	Identifiers required for the support of the domain-specific modelling approach	Author's Experience
HW Device ID		

## 4. APPLICATION

In this section, the HSI specification approach is applied to the development of an automotive fuel tank system for compressed natural gas (CNG). For an appropriate scale of the showcase, only a small part of the real-world system is utilized. The application should be seen as illustrative material, reduced for internal training purposes for students. Therefore, the disclosed and commercially non-sensitivity use-case is not intended to be exhaustive or representative of leading-edge technology. Before the showcase is illustrated, tool support regarding both domain-specific modelling and requirements management shall be explained briefly.

### 4.1. EMS-DSM Language Tool Support

Generally speaking, the EMS-DSM language can be supported by various tools, but at the time when the research project was initiated, the highest possible flexibility was desired, as was full access to the tool's source code. To avoid developing an application from scratch, the open source project *WPF Diagram Designer* (see [18]) was chosen as a basis for tool development. The corresponding documentation has about 540,000 views and the source code has been downloaded more than 24,000 times. Therefore, the source, which provides standard functionality such as file handling and basic graphical modelling, is well reviewed. The source code is written in C# and provides good expandability. New functionalities have been implemented for the diagram designer, named *EASy-Design* (**E**mb**S**ed**A**utomotive **S**ystem-**D**esign), to facilitate engineering with EMS-DSM models. However, EASy-Design is just one possibility for EMS-DSM tool support. The methodology and its C# implementation can be ported to e.g. Enterprise Architect<sup>2</sup> by the provided *Add-in* mechanism. Another alternative is the Eclipse<sup>3</sup> framework, or rather the Eclipse-based project *Sirius*, which enables the creation of a graphical modelling workbench, by facilitating the Eclipse modelling technologies without writing code.

### 4.2. Project and Requirements – Management Tool Support

The web-based open source application *Redmine*<sup>4</sup> is used for topics such as project management and requirements management in this approach. Owing to its high flexibility through configuration, new trackers have been added for development according to the de facto standard Automotive SPICE [4]. The process reference model already mentioned in Section 3 defines three different types of requirements of the engineering process group: *Customer Requirements*, *System Requirements*, and *Software Requirements*. The hardware focus is missing from the embedded E/E system view. Additionally, requirements and design items for mechanical components have to be introduced for the design of an embedded mechatronics-based E/E system. Similar to the Automotive SPICE methodology on a system and software level, engineering processes have been defined for these missing artifacts. To sum up, the available requirement and test case types for this approach are: *Customer Req*, *System Req*, *System TC*, *System Integration TC*, *Software Req*, *Software TC*, *Software Integration TC*, *Hardware Req*, *Hardware TC*, *Mechanics Req*, and *Mechanics TC*.

The test case and requirement items are connected to each other by their unique identifier. For a safety-critical development according to ISO 26262, additional issue types such as *Functional*

---

<sup>2</sup> <http://www.sparxsystems.com/>

<sup>3</sup> <http://eclipse.org/>

<sup>4</sup> <http://www.redmine.org/>

*Safety Requirements* have been added. By reconfiguring the project management tool Redmine, all requirement types mentioned have been implemented.

### 4.3. CNG Tank System Showcase

Figure 5 illustrates the EMS-DSM tool *EASy-Design* with the system architectural design model of the simplified showcase. The CNG fuel tank system consists of seven mechanical components, which are blue coloured (Tank Cylinder, Filter, etc.) The medium flow between mechanical components, which is CNG in this case, is displayed by blue lines with an arrow at the end. Furthermore, five hardware components are placed at the system design model level, which are yellow coloured (In-Tank Temperature Sensor, Tank ECU, etc.) The signal flow between the components is displayed using yellow lines ending with an arrow. A communication bus is inserted between the *Control Unit* and the *External Control Unit* component, shown by the double compound line type and arrows at both ends.

By selecting a model element and clicking the button *Link Requirements*, the elements requirements dialogue is opened and a link between the selected element and an item from the requirements database (e.g. *System Requirement*, see Subsection 4.2) can be established. Already linked requirements from Redmine's MySQL database are listed with their ID, Type, Title, ASIL, and Core functionality attribute. With a click on *Link Specifications*, various documents, such as detailed hardware specifications and datasheets, can be linked to the selected model element.

The *Hardware-Software Interface Specification* emphasis of this contribution is also supported by *EASy-Design*. Again, a hardware element of the model has to be selected and can be defined with a subsequent click on the button *Edit Hardware-Software Interface* in the *Element Properties* group, the interface of the selected hardware item. In Figure 5, the Tank ECU has been selected and in Figure 6, the newly opened HSI definition dialogue for the Tank ECU is illustrated. Within this dialogue, all operations needed to add, modify or delete signals can be triggered by clicking the relevant button:

- *Add New Signal* - a new dialogue window is opened and a signal can be created by entering the properties described in Table 2 (see Figure 7).
- *Add Connected Signal* - the hardware elements in the architectural system design can be connected by (yellow) lines as described in Subsection 4.1. Every output signal from any connected hardware element can be added as an input signal in the HSI signal definition in the actual hardware element.
- *Modify Signal* - at the HSI signal definition main dialogue (illustrated in Figure 6), a signal has to be selected, for which the modification dialogue will be opened after a click on *Modify Signal*. The signal modification dialogue is similar to the *Add New Signal* dialogue.
- *Import Signal(s)* - the *HSI Definition Importer*, as described in Subsection 3.4, is selected, and signals from a HSI signal definition stored in spreadsheet format can be added to the system architectural design model.
- *Export Signal(s)* - the *HSI Definition Exporter*, as described in Subsection 3.4, is selected and signals from the HSI signal definition in the system architectural design model can be exported to a HSI signal definition in spreadsheet format.
- *Delete Signal(s)* - the signals have to be selected from the main HSI signal definition dialogue and are removed from the interface when the button is clicked.

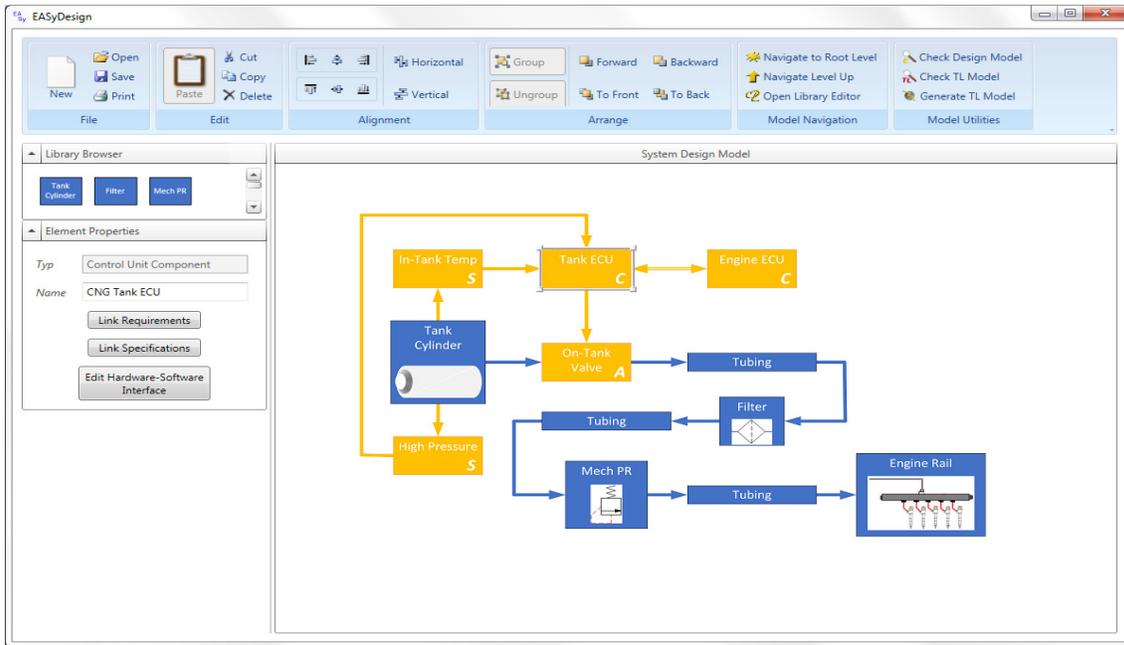


Figure 5. Self-developed tool *EASy Design* with a Simplified CNG Tank System Architectural Design

Direction	HW Signal Name	Description	Sensor / Actuator	Supply Voltage	Physical Min Value	Physical Max Value	Ac
Input	CNGTankT	CNG In-Tank Temperature Signal	Heraeus W-EYK 6 PT100	5V	1V	4V	5%
Input	CNGTankHiP	CNG Tank High Pressure Signal	Sensata Dimensions CNG-PP	5V	0.5V	4.5V	2.5%
Output	CNGOnTankVlvCtrl	Open / Close CNG On-Tank Valve	OMB Lyra224	12V	0V	16V	-
Input	CNGRailT	Gas Temperature at CNG Rail	CAN A / Engine ECU	-	-	-	-
Input	CNGenaSply	Request CNG Supply	CAN A / Engine ECU	-	-	-	-
Input	CNGLoP	CNG Low Pressure at Rail	CAN A / Engine ECU	-	-	-	-
Output	CNGTankFillLvl	CNG Tank Fill Level	CAN A / Engine ECU	-	-	-	-

Figure 6. Hardware-Software Interface Dialog at *EASy Design*

Figure 7. Hardware-Software Interface Add New Signal Dialog at *EASy Design*

As can be seen in Figure 5, no *Software Components* are modelled at this level (System Design Model). With a double-click on a *Hardware Component* (e.g. *Tank ECU*), the next modelling level is opened (named *E/E Item Design Level*). The (green coloured) *Basis Software Components* and *Application Software Components* can be placed here. At each basis software component, the input and output signals from the HSI definition in the particular hardware component can be used and therefore connected to the software.

## 5. CONCLUSION

Previous sections described the factors influencing the development of our hardware-software interface specification approach as well as the supporting tools. A domain-specific modelling method for the design of embedded automotive mechatronics-based E/E systems formed the basis for this work. This approach has the potential to bring together the different engineering disciplines involved in E/E system development by facilitating the HSI specification process. Additionally, many artifacts such as requirements, verification criteria, and various specifications can be linked to the models, created with the new, domain-specific modelling language. With the help of the linked artifacts, vital traceability can be established. Depending on the respective tool chain and the organisation's process landscape, the EMS-DSM models can also facilitate a single point of truth strategy.

First use case implementations show promising results. However, there are several features that still need to be implemented. Options for describing the system's behaviour, e.g. a kind of task scheduling definition, are to be introduced. Furthermore, the Model-to-Model-Transformer between the domain-specific and traditional SysML system architectural design model has to be extended to achieve an automatic transformation of the HSI signal definition between the different modelling strategies.

## REFERENCES

- [1] S. Friedenthal, A. Moore, and R. Steiner, "OMG Systems Modeling Language (OMG SysMLTM) Tutorial," in INCOSE International Symposium, 2006.
- [2] J. Meseguer, "Why Formal Modeling Language Semantics Matters," in Model-Driven Engineering Languages and Systems, ser. Lecture Notes in Computer Science, J. Dingel, W. Schulte, I. Ramos, S. A. ao, and E. Insfran, Eds., vol. 17th International Conference, MODELS 2014, Valencia, Spain, no. 8767. Springer International Publishing Switzerland, 2014, keynote.
- [3] "ISO 26262, Road vehicles - Functional safety," International Organization for Standardization, Geneva, CH, International Standard, November 2011.
- [4] VDA QMC Working Group 13 / Automotive SIG, "Automotive SPICE Process Assessment / Reference Model," Tech. Rep. Revision ID: 470, July 2015, version 3.0.
- [5] E. Andrianarison and J.-D. Piques, "SysML for embedded automotive Systems: a practical approach," in Conference on Embedded Real Time Software and Systems. IEEE, 2010.
- [6] R. Boldt, "Modeling AUTOSAR systems with a UML/SysML profile," IBM Software Group, Tech. Rep., July 2009.
- [7] H. Giese, S. Hildebrandt, and S. Neumann, "Model Synchronization at Work: Keeping SysML and AUTOSAR Models Consistent," LNCS 5765, pp. 555 –579, 2010.

- [8] R. Kawahara, H. Nakamura, D. Dotan, A. Kirshin, T. Sakairi, S. Hirose, K. Ono, and H. Ishikawa, "Verification of embedded system's specification using collaborative simulation of SysML and simulink models," in International Conference on Model Based Systems Engineering (MBSE'09). IEEE, 2009, pp. 21–28.
- [9] J. Meyer, "Eine durchgängige modellbasierte Entwicklungsmethodik für die automobiler Steuergeräteentwicklung unter Einbeziehung des AUTOSAR Standards," Ph.D. dissertation, Universität Paderborn, Fakultät für Elektrotechnik, Informatik und Mathematik, Paderborn, Germany, July 2014.
- [10] M. Broy, M. Feilkas, M. Herrmannsdoerfer, S. Merenda, and D. Ratiu, "Seamless Model-Based Development: From Isolated Tools to Integrated Model Engineering Environments," Proceedings of the IEEE, vol. 98, no. 4, pp. 526–545, 2010.
- [11] M. Mernik, J. Heering, and A. M. Sloane, "When and how to develop domain-specific languages," ACM computing surveys (CSUR), vol. 37, no. 4, pp. 316–344, 2005.
- [12] C. Preschern, N. Kajtazovic, and C. Kreiner, "Efficient development and reuse of domain-specific languages for automation systems," International Journal of Metadata, Semantics and Ontologies, vol. 9, no. 3, pp. 215–226, 2014.
- [13] V. Vujovic, M. Maksimovic, and B. Perisic, "Sirius: A rapid development of DSM graphical editor," in 18th International Conference on Intelligent Engineering Systems (INES). IEEE, 2014, pp. 233–238.
- [14] G. Macher, H. Sporer, E. Armengaud, E. Brenner, and C. Kreiner, "Using Model-based Development for ISO26262 aligned HSI Definition," in Critical Automotive applications: Robustness & Safety, ser. CARS@EDCC2015, Paris, France, 2015.
- [15] H. Sporer, "A Model-Based Domain-Specific Language Approach for the Automotive E/E-System Design," in International Conference on Research in Adaptive and Convergent Systems (RACS 2015), ser. RACS '15, Prague, Czech Republic, 2015.
- [16] G. Macher, E. Armengaud, and C. Kreiner, "Bridging Automotive Systems, Safety and Software Engineering by a Seamless Tool Chain," in 7th European Congress Embedded Real Time Software and Systems Proceedings, 2014, pp. 256–263.
- [17] Software Engineering Institute, "CMMI for Development, Version 1.3," SEI, Carnegie Mellon, Tech. Rep. CMU/SEI-2010-TR-033, ESCTR-2010-033, November 2010.
- [18] Code Project, "WPF Diagram Designer - Part 4," Online Resource, March 2008, <http://www.codeproject.com/Articles/24681/WPFDiagram-Designer-Part>, accessed Mar 2015.

## AUTHORS

**Harald Sporer** received a MSc. degree in Telematics from Graz University of Technology. He worked as software development engineer on Hardware-in-the-Loop (HIL) systems at AVL List GmbH and as functional software developer for embedded automotive systems at Magna Powertrain AG & Co KG. Currently he is working on his PhD at the Institute of Technical Informatics at Graz University of Technology. Parallel to his PhD thesis he is also active in the field of embedded automotive system design, engineering process improvement, and functional safety engineering.



**Georg Macher** received a MSc. degree in Telematics and worked as software development engineer on prototype vehicles at AVL List GmbH. Currently he joined the R&D department of AVL's powertrain engineering branch and is working on his PhD at Institute for Technical Informatics at Graz University of Technology. Parallel to his PhD thesis is also active in the field of system, software, and functional safety engineering.



**Dr. Christian Kreiner** graduated and received a PhD degree in Electrical Engineering from Graz University of Technology in 1991 and 1999 respectively. 1999-2007 he served as head of the R&D department at Salomon Automation, Austria, focusing on software architecture, technologies, and processes for logistics software systems. He was in charge to establish a company-wide software product line development process and headed the product development team. During that time, he led and coordinated a long-term research programme together with the Institute for Technical Informatics of Graz University of Technology. There, he currently leads the Industrial Informatics and Model-based Architectures group. His research interests include systems and software engineering, software technology, and process improvement.



**Prof. Dr. Eugen Brenner** is Associate Professor at the Institute for Technical Informatics of the Graz University of Technology. He completed his master in Electrical Engineering 1983 in Graz. His PhD in Control Theory was finished 1987 also in Graz, dealing with optimal control in systems with limited actuating variables. He joined the institute in 1987, being the first scientific staff member. His post-doctoral lecture qualification in Process Automation was achieved in 1996. He has been member of the senate, of the curricula commission for Bachelor and Master-Programs, and Dean of Studies for Telematics. He currently is head of the Study Commission and Vice-Dean of Studies for Telematics. Eugen Brenner's primary research interests developed from FPGA-based hardware extension to parallel systems, real-time systems and process control systems. The most recent focus targeting embedded systems is on modelling, software-development, systems engineering and systems security, including agile programming methods and smart service engineering.



*INTENTIONAL BLANK*

# NEAR-REAL-TIME PARALLEL ETL+Q FOR AUTOMATIC SCALABILITY IN BIGDATA

Pedro Martins, Maryam Abbasi, Pedro Furtado

Department of Informatics, University of Coimbra, Portugal  
pmom@dei.uc.pt, maryam@dei.uc.pt, pnf@dei.uc.pt

## ABSTRACT

*In this paper we investigate the problem of providing scalability to near-real-time ETL+Q (Extract, transform, load and querying) process of data warehouses. In general, data loading, transformation and integration are heavy tasks that are performed only periodically during small fixed time windows.*

*We propose an approach to enable the automatic scalability and freshness of any data warehouse and ETL+Q process for near-real-time BigData scenarios. A general framework for testing the proposed system was implementing, supporting parallelization solutions for each part of the ETL+Q pipeline. The results show that the proposed system is capable of handling scalability to provide the desired processing speed.*

## KEYWORDS

*Scalability, ETL, freshness, high-rate, performance, parallel processing, distributed systems, database, load-balance, algorithm*

## 1. INTRODUCTION

ETL tools are special purpose software used to populate a data warehouse with up-to-date, clean records from one or more sources. The majority of current ETL tools organize such operations as a workflow. At the logical level, the E (Extract) can be considered as a capture of data-flow from the sources with more than one high-rate throughput. T (Transform) represents transforming and cleansing data in order to be distributed and replicated across many processes and ultimately, L(Load) the data into data warehouses to be stored and queried. For implementing these type of systems besides knowing all of these steps, the acknowledgement of user regarding the scalability issues is essential.

When defining the ETL+Q the user must consider the existence of data sources, where and how the data is extracted to be transformed, loading into the data warehouse and finally the data warehouse schema; each of these steps requires different processing capacity, resources and data treatment. Moreover, the ETL is never so linear and it is more complex than it seems. Most often the data volume is too large and one single extraction node is not sufficient. Thus, more nodes must be added to extract the data and extraction policies from the sources such as round-robin or on-demand are necessary.

After extraction, data must be re-directed and distributed across the available transformation nodes, again since transformation involves heavy duty tasks (heavier than extraction), more than one node should be present to assure acceptable execution/transformation times. Once again more new data distribution policies must be added. After the data is transformed and ready to be loaded, the load period time and a load time control must be scheduled. This means that the data have to be held between the transformation and loading process in some buffer. Eventually, regarding the data warehouse schema, the entire data will not fit into a single node, and if it fits, it will not be possible to execute queries within acceptable time ranges. Thus, more than one data warehouse node is necessary with a specific schema which allows to distribute, replicate, and query the data within an acceptable time frame.

In this paper we study how to provide parallel ETL+Q scalability with ingress high-data-rate in big data warehouses. We propose a set of mechanisms and algorithms to parallelize and scale each part of the entire ETL+Q process, which later will be included in an auto-scale (in and out) ETL+Q framework. The presented results prove that the proposed mechanisms are able to scale when necessary.

In Section 2 we present some relevant related-work in the field. Section 3 describes the architecture of the proposed system, Section 4 explains the main algorithms which allow to scale-out when necessary. Section 5 shows the experimental results obtained when testing the proposed system. Finally, Section 6 concludes the paper and discusses future work.

## 2. RELATED WORK

Works in the area of ETL scheduling includes efforts towards the optimization of the entire ETL workflows [8] and of individual operators in terms of algebraic optimization; e.g., joins or data sort operations.

The work [4] deals with the problem of scheduling ETL workflows at the data level and in particular scheduling protocols and software architecture for an ETL engine in order to minimize the execution time and the allocated memory needed for a given ETL workflow. The second aspect in ETL execution that the authors address is how to schedule flow execution at the operations level (blocking, non-parallelizable operations may exist in the flow) and how we can improve this with pipeline parallelization [3].

The work [6] focuses on finding approaches for the automatic code generation of ETL processes which is aligning the modeling of ETL processes in data warehouse with MDA (Model Driven Architecture) by formally defining a set of QVT (Query, View, Transformation) transformations.

ETLMR [5] is an academic tool which builds the ETL processes on top of Map-Reduce to parallelize the ETL operation on commodity computers. ETLMR does not have its own data storage (note that the offline dimension store is only for speedup purpose), but is an ETL tool suitable for processing large scale data in parallel. ETLMR provides a configuration file to declare dimensions, facts, User Defined Functions (UDFs), and other run-time parameters.

In [7] the authors consider the problem of dataflow partitioning for achieving real-time ETL. The approach makes choices based on a variety of trade-offs, such as freshness, recoverability and

fault-tolerance, by considering various techniques. In this approach partitioning can be based on round-robin (RR), hash (HS), range (RG), random, modulus, copy, and others [9].

In [2] the authors describe Liquid, a data integration stack that provides low latency data access to support near real-time in addition to batch applications. It supports incremental processing, and is cost-efficient and highly available. Liquid has two layers: a processing layer based on a statefull stream processing model, and a messaging layer with a highly-available publish / subscribe system. The processing layer (i) executes ETL-like jobs for different back-end systems according to a stateful stream processing model [1]; (ii) guarantees service levels through resource isolation; (iii) provides low latency results; and (iv) enables incremental data processing. A messaging layer supports the processing layer. It (i) stores high-volume data with high availability; and (ii) offers rewindability, i.e. the ability to access data through meta-data annotations. The two layers communicate by writing and reading data to and from two types of feeds, stored in the messaging layer.

Related problems studied in the past include the scheduling of concurrent updates and queries in real-time warehousing and the scheduling of operators in data streams management systems. However, we argue that a fresher look is needed in the context of ETL technology. The issue is no longer the scalability cost/price, but rather the complexity it adds to the system. Previews presented recent works in the field do not address in detail how to scale each part of the ETL+Q and do not regard the automatic scalability to make ETL scalability easy and automatic. The authors focus on mechanisms to improve scheduling algorithms and optimizing work-flows and memory usage. In our work we assume that scalability in number of machines and quantity of memory is not the issue, we focus on offering scalability for each part of the ETL pipeline process, without the nightmare of operators relocation and complex execution plans. Our main focus is on scalability based on generic ETL process to provide the users desired performance with minimum complexity and implementations. In addition, we also support queries execution.

### 3. ARCHITECTURE

In this section we describe the main components of the proposed architecture for ETL+Q scalability. Figure 1 shows the main components to achieve automatic scalability.

- All components from (1) to (7) are part of the Extract, Transform, Load and query (ETL+Q) process.
- The "Automatic Scaler" (13), is the node responsible for performance monitoring and scaling the system when it is necessary.
- The "Configuration file" (12) represents the location where all user configurations are defined by the user.
- The "Universal Data Warehouse Manager" (11), based on the configurations provided by the user and using the available "Configurations API" (10), sets the system to perform according with the desired parameters and algorithms. The "Universal Data Warehouse Manager" (11), also sets the configuration parameters for automatic scalability at (13) and the policies to be applied by the "Scheduler" (14).

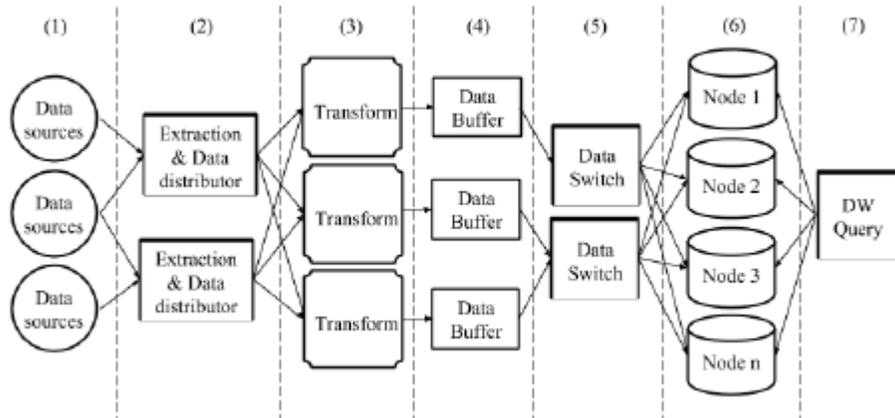


Figure 2: Total automatic ETL+Q scalability

- The "Configuration API" (10), is an access interface which allows to configure each part of the proposed Universal Data Warehouse architecture, automatically or manually by the user.
- Finally the "Scheduler" (14), is responsible for applying the data transfer policies between components (e.g. control the on-demand data transfers).

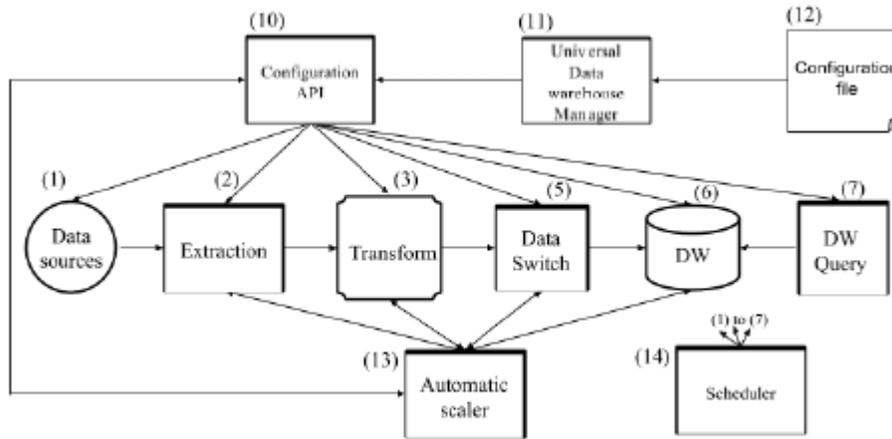


Figure 1: Automatic ETL+Q scalability

All these components when set to interact together are able to provide automatic scalability to the ETL and to the data warehouses processes without the need for the user to concern about its scalability or management.

**Paralelization approach**

Figure 2 depicts the main processes needed to support total ETL+Q scalability.

- 1) Represents the data sources from where data is extracted into the system.

- 2) The data distributor(s) is responsible for forwarding or replicating the raw data to the transformer nodes. The distribution algorithm to be used is configured and enforced in this stage. The data distributors (2) should also be parallelizable if needed, for scalability reasons.
- 3) In the transformation nodes the data is cleaned and transformed to be loaded into the data warehouse. This might involve data lookups to in-memory or disk tables and further computation tasks. The transformation is parallelized for scalability reasons.
- 4) The data buffer can be in memory, disk file (batch files) or both. In periodically configured time frames/periods, data is distributed across the data warehouse nodes.
- 5) The data switches are responsible to distribute (pop/extract) data from the "Data Buffers" and set it for load into the data warehouse, which can be a single-node or a parallel data warehouse depending on configured parameters (e.g. load time, query performance).
- 6) The data warehouse can be in a single node, or parallelized by many nodes. If it is parallelized, the "Data Switch" nodes will manage data placement according to configurations (e.g. replication and distribution). Each node of the data warehouse loads the data independently from the batch files.
- 7) Queries are rewritten and submitted to the data warehouse nodes for computation. The results are then merged, computed and returned.

The main concept we propose are the individual ETL+Q scalability mechanisms of each part of the ETL+Q pipeline. By offering solution to scale each part independently, we provide a solution to obtain configurable performance.

#### **4. DECISION ALGORITHMS FOR SCALABILITY PARAMETERS**

In this section we define the scalability decision methods as well as the algorithms which allow the framework to automatically scale-out and scale-in.

##### **Extraction & data distributors - Scale-out**

Depending on the number of existing sources and data generation rate and size, the nodes that process data extraction from the sources might need to scale. The addition of more "extraction & data distributors" (2) depends if the current number of nodes is being able to extract and process the data with the correct frequency (e.g. every 5 minutes) and inside the limit maximum extraction time (without delays). For instance, if the extraction frequency is specified as every 5 minutes and extraction duration 10 seconds, every 5 minutes then the "Extraction & Data distributor" nodes cannot spend more than 10 seconds extracting data. Otherwise a scale-out is needed, so the extraction size can be reduced and the extraction time improved. If the maximum extraction duration is not configured, then the extraction process must finish before the next extraction instant. If not processed until the next extraction instant, as defined by the extraction frequency, a scale-out is also required, to add more extraction power. Flowchart 3 describes the algorithm used to scale-out.

**Extraction & data distributors- Scale-in**

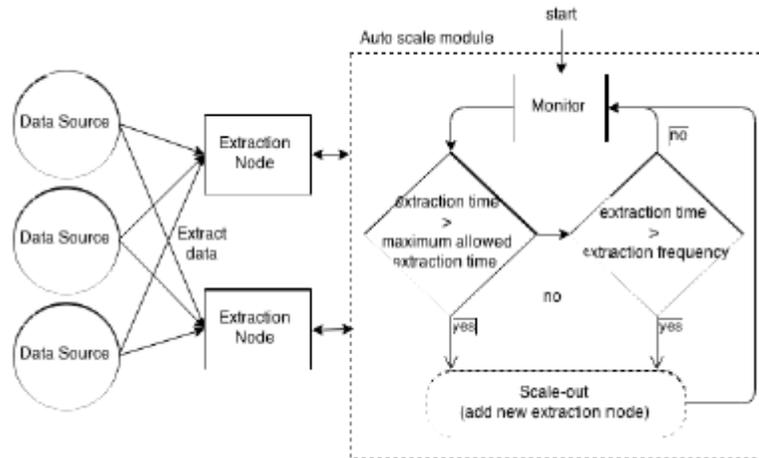


Figure 3: Extraction algorithm-scale-out

To save resources when possible, nodes that perform the data extraction from the sources can be set in standby or removed. This decision is made based on the last execution times. If previous execution times of at least two or more nodes are less than half of the maximum extraction time (or if the maximum extraction time is not configured, the frequency), minus a configured variation parameter (X), one of the nodes is set on standby or removed, and the other one takes over. Flowchart 4 describes the applied algorithm to scale-in.

**Transform - Scale-out**

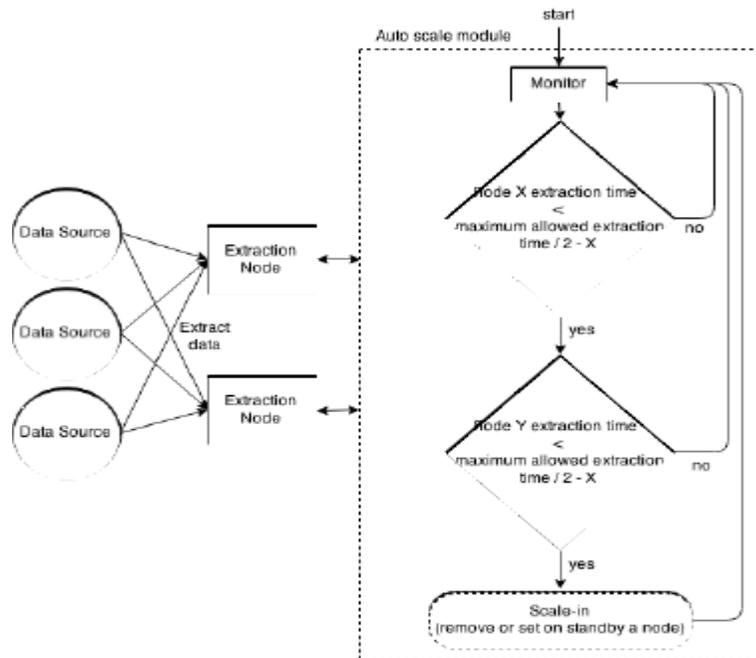


Figure 3: Extraction algorithm-scale-out

The transformation process is critical, if the transformation is running slow, data extraction at the referred rate may not be possible, and information will not be available for loading and querying when necessary. The transformation step has an important queue used to determine when to scale the transformation phase. If this queue reaches a limit size (by default 50%), then it is necessary to scale, because the actual transformer node(s) is not being able to process all the data that is arriving. The size of all queues is analyzed periodically. If this size at a specific moment is less than half of the limit size for at least two nodes, then one of those nodes is set on standby or removed. Flowchart 5, describes the algorithm used to scale-out and scale-in.

### Data buffer - Scale

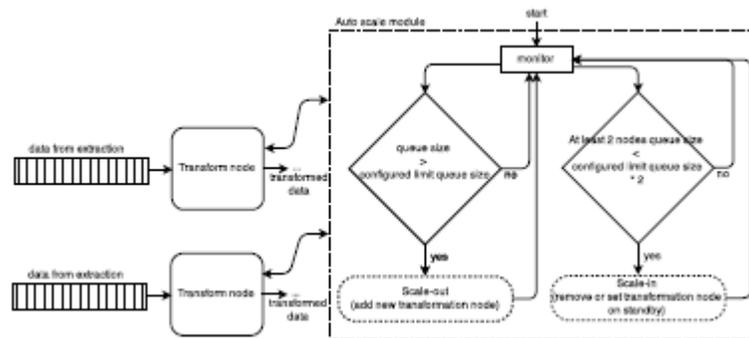


Figure 5: Transformation - scale-in and scale-out

The data buffer nodes scale-out based on the incoming memory queue size and the storage space available to hold data. When the the available memory queue becomes full, above 50% of the configured maximum size, data starts being swapped into disk, until the memory is empty. If even so the data buffer memory reaches the limit size the data buffer must be scaled-out. This means that the incoming data-rate (going into memory storage) is not fast enough to swap to the disk storage and more nodes are necessary. If the disk space becomes full above a certain configured size, the data buffers are also set to scale-out. Flowchart 6 describes the algorithm used to scale-out the data buffer nodes. By user request the data buffers can also scale-in, in this case the system will scale if the data from any data buffer can be fitted inside another data buffer.

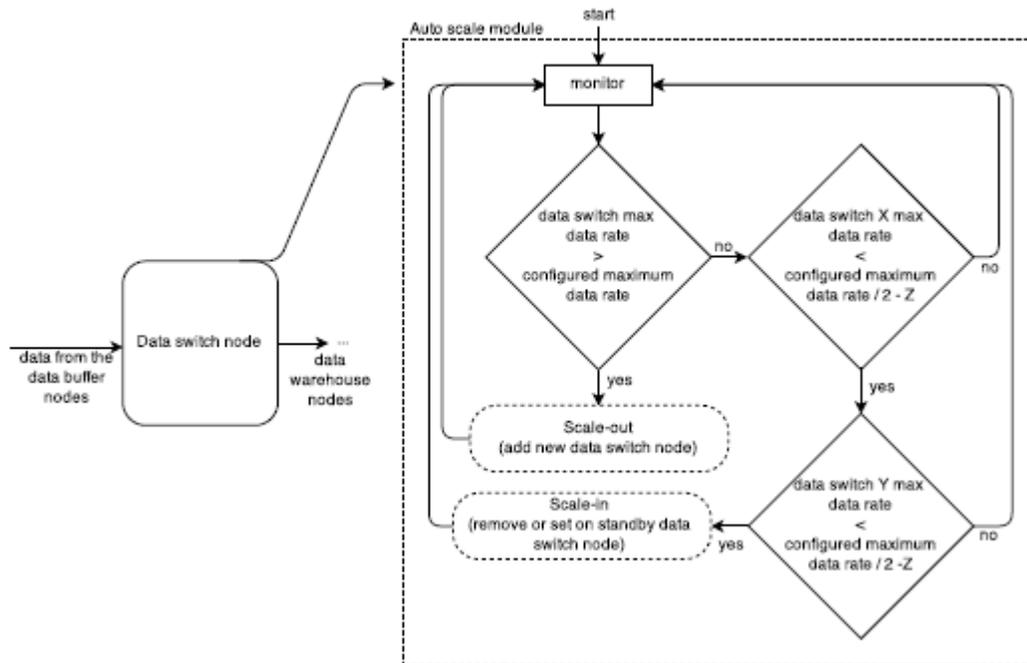


Figure 7: Data switch – scale

**Data switch – Scale**

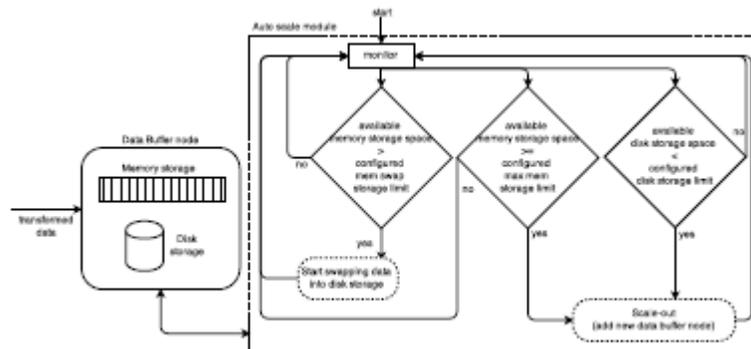


Figure 6: Data Buffers-scale-out

The Data Switch nodes scale based on a configured data-rate limit. If the data-rate rises above the configured limit the data switch nodes are set to scale-out. The data switches can also scale-in, in this case the system will allow it if the data-rate is less than the configured maximum by at least 2 nodes, minus a Z configured variation, for a specific time. Flowchart 7 describes the used algorithm to scale the data switch nodes.

**Data Warehouse – Scale**

Data warehouse scalability needs are detected after each load process or by query execution time. The data warehouse load process has a configured limit time to be executed every time it starts. If that limit time is exceeded, then the data warehouse must scale-out. Flowchart 8 describes the algorithm used to scale the data warehouse when the maximum load time is exceeded.

The data warehouse scalability is not only based on the load & integration speed requirements, but also on the queries desired maximum execution time. After each query execution, if the query time to the data warehouse is more than the configured maximum desired query execution time, then the data warehouse is set to scale-out. The Flowchart 9 describes the algorithm used to scale the data warehouse based on the average query execution time.

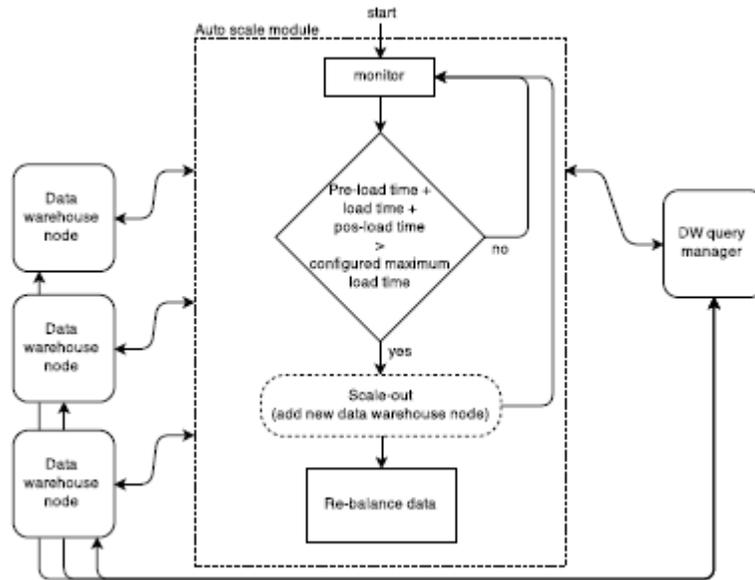


Figure 8: Data warehouse – scale

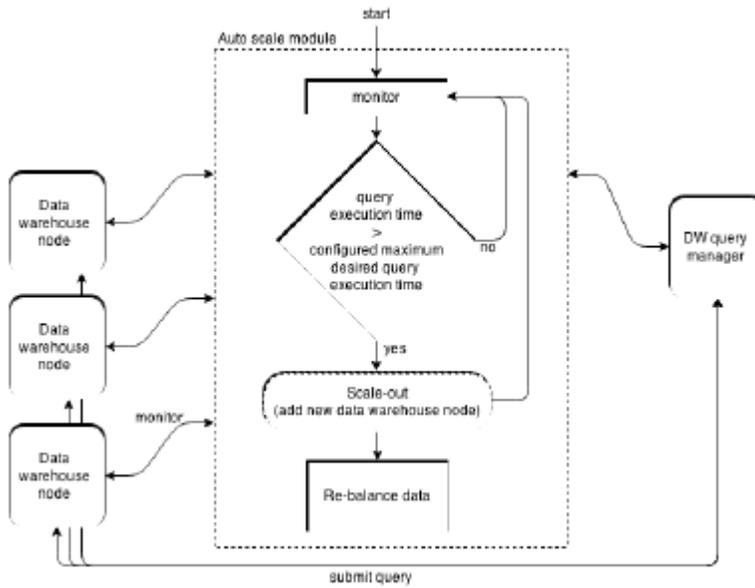


Figure 9: Data warehouse - scale based on query time

The data warehouse nodes scale-in is performed if the average query execution time and the average load time respect the conditions 1 and 2 (where  $n$  represents the number of nodes):

$$\frac{(n-1) \times avgQueryTime}{n} \leq desiredQueryTime \quad (1)$$

and

$$\frac{(n-1) \times avgLoadTime}{n} \leq maxLoadTime \quad (2)$$

Every time the data warehouse scales-out or scales-in the data inside the nodes needs to be re-balanced. The default re-balance process to scale-out is based on the phases: Replicate dimension tables; Extract information from nodes; Load the extracted information into the new nodes. Scale-in process is more simple, data just needs to be extracted and loaded across the available nodes as if it is new data.

## 5. EXPERIMENTAL SETUP AND RESULTS

In this section we describe the experimental setup, and experimental results to show that the proposed system, AScale, is able to scale and load balance data and processing.

### 5.1 Experimental Setup

In this section we describe the used testbed. The experimental tests were performed using 12 computers, denominated as nodes, with the following characteristics: Processor Intel Core i5-5300U Processor (3M Cache, upto 3.40 GHz); Memory 16GB DDR3; Disk: western digital 1TB 7500rpm; Ethernet connection 1Gbit/sec; Connection switch: SMC SMCOST16, 16 Ethernet ports, 1Gbit/sec.

The 12 nodes were formatted before the experimental evaluation and installed with: Windows 7 enterprise edition 64 bits; Java JDK 8; Netbeans 8.0.2; Oracle Database 11g Release 1 for Microsoft Windows (X64) - used in each data warehouse nodes; PostgreSQL 9.4 - used for look ups during the transformation process; TPC-H benchmark - representing the operational log data used at the extraction nodes. This is possible since TPC-H data is still normalized; SSB benchmark - representing the data warehouse. The SSB is the star-schema representation of TPC-H data.

### 5.2 Automatic scalability

In this section we describe the scalability tests made to the full auto-scale ETL framework. We demonstrate scale-out and scale-in ability of the proposed framework using the proposed algorithms. Consider the following scenario:

- Sources are based on the TPC-H benchmark generator, which generate data at the highest possible rate (on each node);

- The transformation process consists of transforming the TPC-H relational model into a star-schema model, which is the SSB benchmark. This process involves heavy computational memory, temporary storage of large amounts of data, look-ups and data transformations to assure consistency;
- The data warehouse tables schema consist on the same schema from SSB benchmark. Replication and partitioning is assured by AScale, whereas, dimension tables are replicated and fact tables are partitioned across the data warehouse nodes.
- The E, T and L were set to perform every 2 seconds, and cannot last more than 1 second. Thus the ETL process will last at the worst case 3 seconds total;
- The load process was made in batches of 100MB maximum;
- All default configurations of other components were set to use the default AScale configurations.

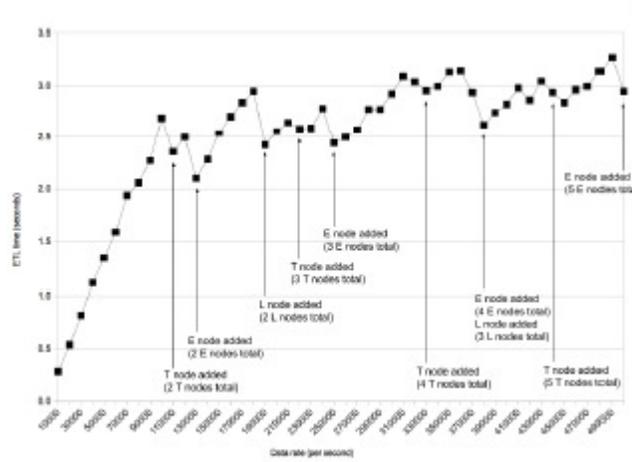


Figure 10: Full ETL system scale-out

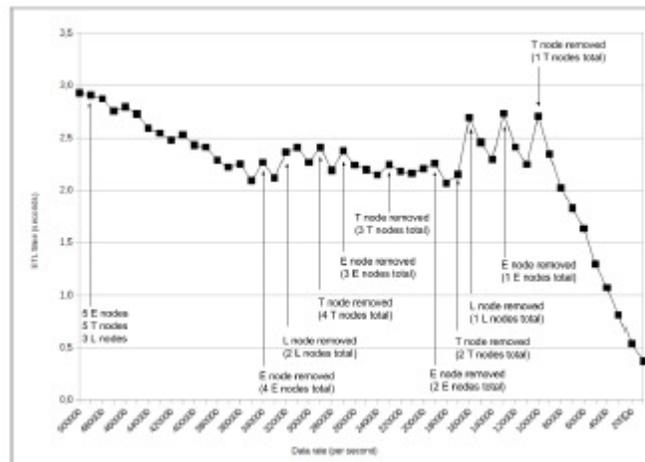


Figure 11: Full ETL system scale-in

Figure 10 and 11 show how the proposed auto-scale ETL framework scales to deliver the configured ETL execution time, while the data rate increases/decreases. In the charts the X axis is represented the data-rate per second, from 10.000 to 500.000 rows per second and the Y axis is the ETL time expressed in seconds; The system objective was set to deliver the ETL process in 3 seconds; In the charts we also represent the scale-out and scale-in of each part of the framework, by adding and removing nodes when necessary. Note that If we set a lower execution time the framework will scale-out faster. If the execution time is higher (e.g. 3 minutes) the framework will scale later if more performance is necessary at any module.

**Scale-out** results from Figure 10 show that, as the data-rate increases and parts of the ETL pipeline become overloaded, by using all proposed monitoring mechanisms in each part of the AScale framework, each individual module scales to offer more performance where necessary. In Figure 10, we point each scaled-out module.

Note that in some stages of the tests the 3 seconds limit execution time was exceeded in 0.1, 0.2 seconds. This happened due to the high data-rate usage of the network connecting all nodes that is not being accounted for the purposes of our tests.

**Scale-in** results from Figure 11 show the moment when the current number of nodes are no longer necessary to assure the desired performance, and some nodes could be removed to be set as ready node in stand-by and be used in other parts to assure the ETL configured time.

When comparing the moments of scale-out and scale-in, it is possible to observe that the proposed framework scales-out much faster than it scales-in. When a scale-in can be applied it is performed in a later stage than a scale-out.

### 5.3 Data extraction nodes scalability

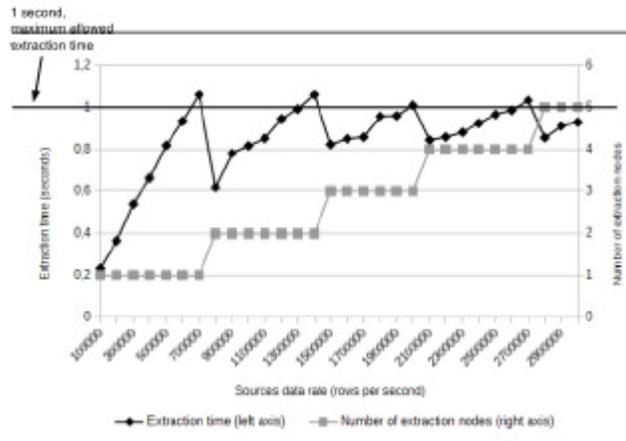


Figure 13: Extraction scalability

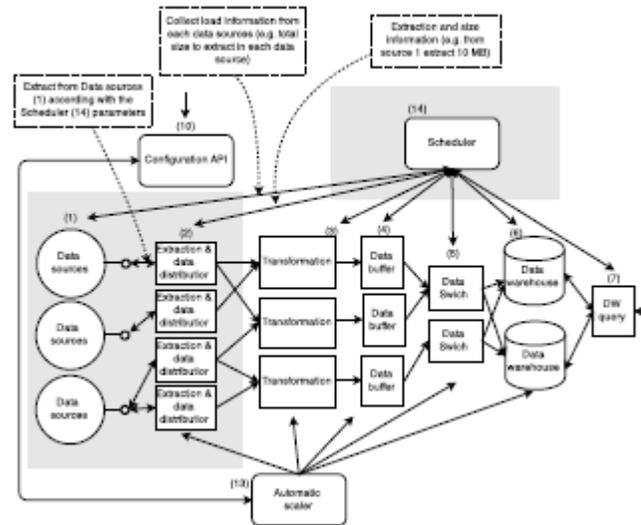


Figure 12: Sources to extraction

Figure 12 shows the scheduler based extraction approach to extract data, where the "automatic scaler" (13) orders the nodes to extract data from sources (scheduler based extraction policy). Considering "data sources" (1) generate high rate data and "extraction nodes" (2) extract the generated data, when the dataflow is too high a single data node cannot handle all ingress data. In this section we study how the extraction nodes scale to handle different data-rates. The maximum allowed extraction time was set to 1 second. Extraction frequency was set to every 3 seconds.

In Figure 13 we have in the left Y axis, the average extraction time in seconds; in the right Y axis, the number of nodes; The X axis is the datarate; Black line represents the extraction time; Grey line represents the number of nodes as they scale-out. It is possible to see that every time the extraction takes too much time (more than 1 second as configured) a new node is added (from the ready-nodes pool). After the new node added, more nodes are being used to extract data from the same number of sources, so the extraction time improves. As we increase the data-rate, the extraction time becomes higher until it reaches more than 1 second, and another node is added.

## 5.4 Transformation scalability

During the ETL process, after data is extracted, it is set for transformation. Because this process is computationally heavy, it is necessary to scale the transformation nodes to assure that all data is processed without delays. Each transformation node has an entrance data queue, for ingress data. The "automatic scaler" (13) monitors all queues, once it detects that a queue is full and above a certain configured threshold it starts the scale process, this means that  $Rate_{extract} \geq Rate_{transform}$

The transformation nodes scale-out mechanisms were set to the limit queue size to trigger the scale-out mechanisms at 50MB, approximately 380.000 rows.

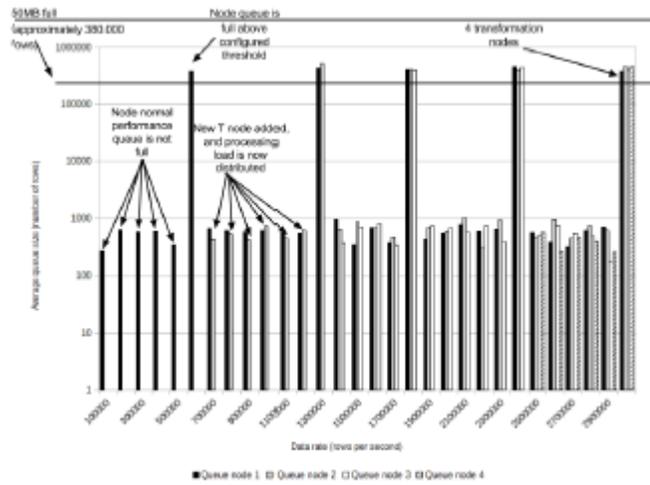


Figure 14: Automatic transformation scalability. 60 minutes processing per data-rate.

In Figure 14 you can see in Y axis, the average queue size in number of rows; in X axis, the data rate in rows per second; Each plotted bar represents the average transformation node queue size (up to 4 nodes); Each measure represents the average queue size of 60 seconds run. As it displays the moments when the queue size of the transformation nodes increase above the configured limit size, a new transformation node is added and data is distributed by all nodes to support the increasing data-rate.

### 5.5 Data Buffer nodes

This nodes hold the transformed data until it is loading into the data warehouse. The data buffers have the following configuration: Generation data rate speed 350.000 rows per second (i.e. transformation output data rate); Available memory storage 10.000MB (50% = 5.000MB); Available disk storage 1TB. We consider only the data generation/producer, there is no data "consumer", so the buffer must hold all ingress data.

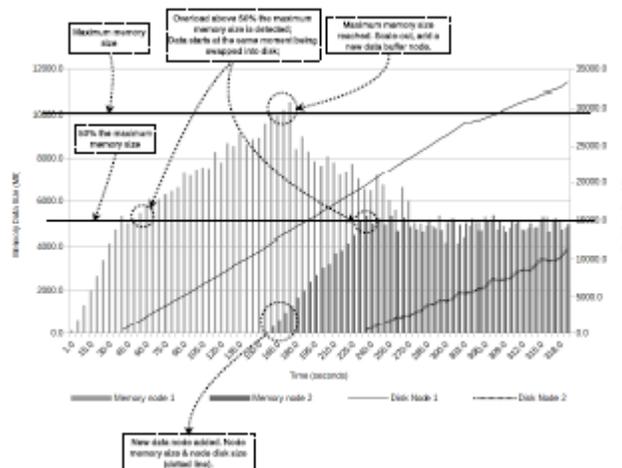


Figure 15: Data buffer swap into disk and scaling

Figure 15 shows: X axis represents the time; Left Y axis represents the memory size; Right Y axis represents the disk storage size; The bars represent the memory usage; Continuous line represent the total data size in disk, being swapped from memory. The chart shows the buffer nodes scalingout. When 50% of the memory size is used, data starts being swapped into disk. However, if the disk cannot handle all ingress data and the memory reaches the maximum limit size, a new node is added. After a new node added, data is distributed by both nodes. This makes the data-rate at each node less (at least half) and then the disk from the first node can empty the memory back to 50%. After this point each time 50% of the maximum configured memory is reached, data will swap into disk to free the memory.

### 5.6 Data warehouse scalability

In this section we test the data warehouse scalability, which can be triggered either by the load process (because it is taking too long), or because the query execution is taking more time than the desired response time.

To test the data warehouse nodes load scalability we set the load method by using batch files of maximum 100Mb. The maximum allowed load time was set to 60 seconds. Each time a data warehouse node is added, we show the data size that was moved into the new node and the required time in seconds to re-balance the data. All load and re-balance times include the execution of Pre-load tasks (i.e. destroy all indexes and views) and Pos-load tasks (i.e. rebuild indexes and update views). If the maximum configured load time is exceeded more than 60 seconds, the data warehouse is set to be scaled.

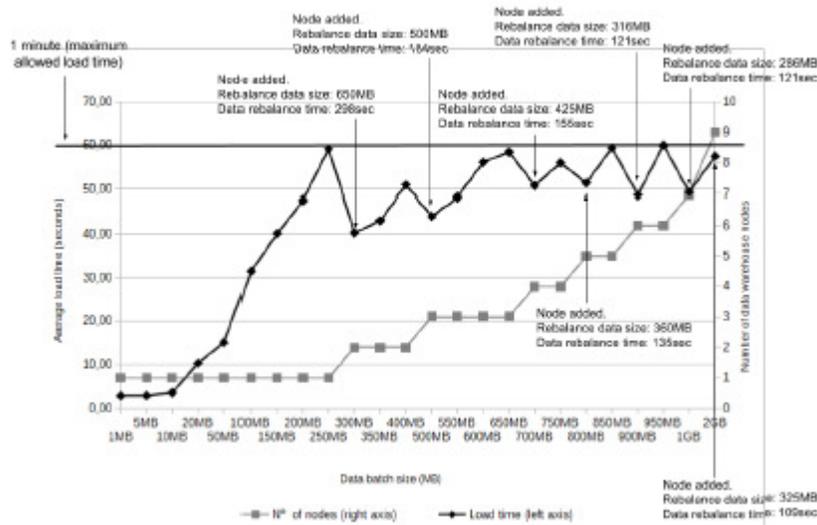


Figure 16: Data warehouse load scalability

In Figure 16 we have in left Y axis, the average load time in seconds; right Y axis, the number of data warehouse nodes; X axis, the data batch size in MB. Horizontal bar at Y = 60 seconds represents the maximum configured load time. At each scale-out moment there is a note specifying the data re-balanced size and time to perform it. Black plotted line represents the average load time. The Grey plotted line represents the number of data warehouse nodes.

Experimental results in this chart shows how the load performance degrades as the data size increases and how it improves when a new node is added. After a new node added performance improves below the maximum configured limit. Note that every time a new node was added, the data warehouse required to be re-balanced (data was evenly distributed by the nodes).

## 5.7 Query scalability

When running queries, if the maximum desired query execution time (i.e. configured parameter) is exceeded, then the data warehouse is set to scale in order to offer more query execution performance. The following workloads were considered to test AScale query scalability, Workload 1 with 50GB total size, executing queries Q1.1, Q2.1, Q3.1, Q4.1 chosen randomly and with a desired execution time per query of 1 minute. Workload 2 with the same properties as workload 1 but, using 1 to 8 simultaneous sessions.

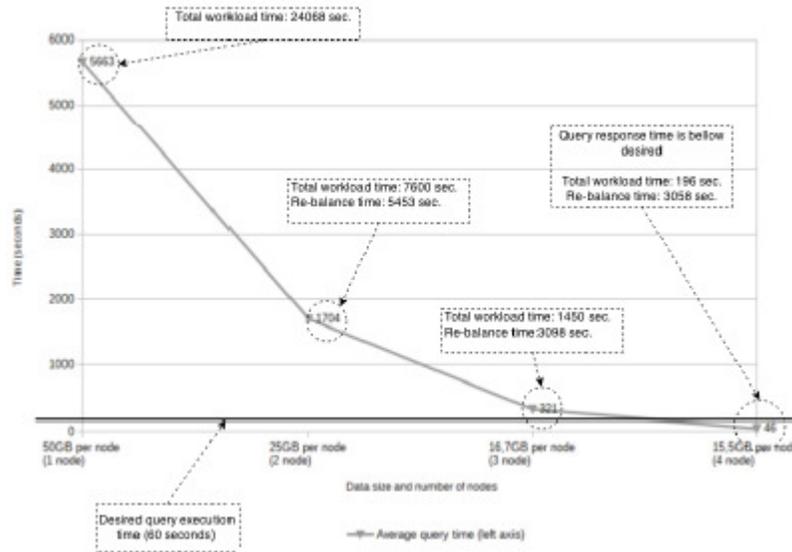


Figure 17: Data warehouse scalability, workload 1

Workload 1 studies how the proposed mechanisms scales-out the data warehouse when running queries. Workload 2 studies the scalability of the system when running queries and the number of simultaneous sessions (e.g. number of simultaneous users) increases. Both workloads were set with the objective of guaranteeing the maximum execution time per query of 60 seconds.

### Query scalability - Workload 1

Figure 17 shows the experimental results for workload 1. Y axis shows the average execution time in seconds and X axis the data size per node and the current number of nodes. The horizontal line over 60 seconds represents the desired query execution time. At each scale-out we identify the total workload time and data re-balance time (i.e. extract data, load into nodes, rebuild indexes and views). The plotted line represents the average query time execution. Every time the average query time is not inferior to the maximum configured query execution time, one extra node is added. In each scale-out, the re-balance time represents the necessary time to extract data from nodes and distribute it across all nodes (we also include indexes and views update

time). Once the average query time reaches under the configured desired execution time, the framework stops scaling the data warehouse nodes.

## Query scalability - Workload 2

Figure 18 shows the experimental results for workload 2. Total data was 50GB. Left Y axis shows the average query execution time in seconds and right Y axis, the average data re-balance time in seconds (i.e. extract from nodes, load into new node, rebuild indexes and views). X axis shows the number of sessions, the data size per node and the number of nodes. The horizontal line over 60 seconds represents the desired query execution time. The last result does not respect the desired execution time because of the limited hardware resources for our tests, 12 nodes. The results show that while the number of simultaneous sessions increases the system scales the number of nodes in order to provide more performance. The query average execution time follows the configured parameters. We also plot the data rebalance time. Every time a new node is added data must be balanced by all data warehouse nodes, this includes, extract data from the existent nodes, load into the new node, and finally re-create all indexes and views (since indexes and views updates are done in parallel for all nodes, we update all indexes and views in the data warehouse simultaneously).

## 6. CONCLUSION & FUTURE WORK

In this work we propose mechanisms and algorithms to achieve automatic scalability for complex ETL+Q, offering the possibility to the users to think solely in the conceptual ETL+Q models and implementations for a single server.

The tests demonstrate that the proposed techniques are able to scale-out automatically when more resources are required. Future work includes the comparison with other state-of-the-art tools and the development of drag and drop interface to make AScale available to public.

## REFERENCES

- [1] R. Castro Fernandez, M. Migliavacca, E. Kalyvianaki, and P. Pietzuch. Integrating scale out and fault tolerance in stream processing using operator state management. In Proceedings of the 2013 ACM SIGMOD international conference on Management of data, pages 725{736. ACM, 2013.
- [2] R. C. Fernandez, P. Pietzuch, J. Koshy, J. Kreps, D. Lin, N. Narkhede, J. Rao, C. Riccomini, and G.Wang. Liquid: Unifying nearline and offline big data integration. In Biennial Conference on Innovative Data Systems Research (CIDR), Asilomar, CA, USA, 01/2015 2015. ACM, ACM.
- [3] R. Halasipuram, P. M. Deshpande, and S. Padmanabhan. Determining essential statistics for cost based optimization of an etl workflow. In EDBT, pages 307{318, 2014.
- [4] A. Karagiannis, P. Vassiliadis, and A. Simitsis. Scheduling strategies for efficient etl execution. Information Systems, 38(6):927{945, 2013.
- [5] X. Liu. Data warehousing technologies for large-scale and right-time data. PhD thesis, dissertation, Faculty of Engineering and Science at Aalborg University, Denmark, 2012.

- [6] L. Munoz, J.-N. Mazon, and J. Trujillo. Automatic generation of etl processes from conceptual models. In Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP, pages 33{40. ACM, 2009.
- [7] A. Simitsis, C. Gupta, S. Wang, and U. Dayal. Partitioning real-time etl workflows, 2010.
- [8] A. Simitsis, K. Wilkinson, U. Dayal, and M. Castellanos. Optimizing etl workflows for fault-tolerance. In Data Engineering (ICDE), 2010 IEEE 26th International Conference on, pages 385{396. IEEE, 2010.
- [9] P. Vassiliadis and A. Simitsis. Near real time etl. In New Trends in Data Warehousing and Data Analysis, pages 1{31. Springer, 2009.

# THE PERCEPTIONS OF AGILE METHODOLOGY IN SOUTH AFRICA

Thierry Mbah Mbelli<sup>1</sup> and Jainesh Jaintylal Hira<sup>2</sup>

<sup>1</sup>Deloitte Digital, Woodmead, Johannesburg, South Africa  
tmbelli@deloitte.co.za

<sup>2</sup>First National Bank, Sandton, Johannesburg, South Africa  
jainesh.hira@fnb.co.za

## ABSTRACT

*Agile methodology was introduced in the mid 90's while the agile manifesto was adopted in 2001. The rationale behind the introduction of the agile methodology was to uncover better ways of developing software that will meet the user's expectation in an iterative controlled manner. With technological explosion and rift competition for market share, user experience and satisfaction can only be achieved through proper communication between stakeholders and innovative ways of doing things. Doing things differently is what the agile methodology brought. Despite the existence of this methodology for over 20 years now, South African software industry is only starting to realize its existence with a lot of companies jumping into the bandwagon. This paper presents the results of an empirical research of how the South African software industry perceive the methodology.*

## KEYWORDS

*Agile methodology, Software crisis, Agile manifesto, Scrum, Extreme programming*

## 1. INTRODUCTION

Every business and every economy depends on technology to survive in a very competitive space nowadays and this has created what some people called 'technology war' or 'software war'. Agile methodology helps to cushion this war and prepares the software industry to manoeuvre the war by introducing methods that place more emphasis on people and their creativity, communication and the ability to produce quality software within a relatively short space of time and within budget. Agile methodology is a very popular approach to developing software but very little is understood about its penetration, successes, benefits, failures and problems [5]. Scrum and Extreme Programming are the most popular agile methodologies that are currently used in South Africa

This paper presents an understanding of how the agile methodology is being perceived in South Africa, what is the general understanding of the methodology by the industry practitioners. Empirical research method was used to gain this understanding. Through this understanding, an insight of the interaction of the development practices and the perceptions of the agile methodology was established.

A web-based survey targeting software development professionals involved in the development, testing and management was conducted. Invitations were sent out by email to 905 selected professionals with instructions of how to complete the anonymous web-based questionnaires. The questions were structured to understand the respondents' understanding of the agile methodology, which agile method was adopted in their team, the successes and failures of agile methodology, their experience in software development, their role in the team, their qualification and their perception of why the methodology works or not work in their development teams. The response rate was 20% with 181 responses received.

From the responses, a perception of the agile methodology was established and scrum was identified as the most common of the agile methods adopted. 58% of the respondents have at least two years' experience in the development of software. 40.9% have at least an undergraduate degree. Less than 8.8% have a postgraduate degree. About 41.4% indicated that agile methodology doesn't work in their teams, while 44.2% indicated that agile methodology is working in their teams, thus a slight approval of the methodology.

The remainder of this paper is structured as follow: Section 2 discusses what work has been done in different parts of the world on this topic. Section 3 discusses the methodology that was used to come out with the findings. Section 4 presents the findings and lastly, section 5 concludes with a review of the findings and the implications for future research in the South African context.

## **2. LITERATURE REVIEW**

Agile methodology represents a major shift from the traditional approach of developing software to a more engineering-like approach. Prior to the emergence of the agile methodology, software development had issues and problems – 'software crisis'. The software crisis was due to the complexity of problems that the engineering methods available could not tackle [3]. The agile methodology brought some shift in the approach to solve complex problems and the adoption of this methodology depends on the suitability and skills of the team members, the organization and the type of project.

The agile manifesto prophesizes four main value [2]:

- More interaction between team members and less emphasis on processes and tools
- More emphasis on producing a working software and less emphasis on detailed documentation
- More customer collaboration in the development process and less contract negotiation
- Ability to rapidly respond to change than following a detailed plan

There is a general perception that agile methodology is widely used in South Africa but there is little or no evidence to validate this perception. This paper aims at addressing this perception and to some level assesses the current state of the practice of the methodology in South Africa.

This research is closely related to the work of Begel and Nagappan [1]. They investigated the usage and perception of agile software development in Europe, Asia and North America

specifically at Microsoft. Africa in general and South Africa in particular was not included in that investigation. They conducted a web based survey of Microsoft employees in the development, testing and management teams. Their research was based at understanding the respondents' demographics, usage and penetration of agile development practices and the general perception of the agile methods. They found out that overwhelming majority of the respondents favoured the adoption of the agile methods.

Nithila et al.'s [4] also conducted a similar study in Sri Lanka and they found out that majority of the respondents favoured the adoption of the agile methodology. In their study they found out that only 31% of the respondents disapproved the agile methodology.

### **3. METHODOLOGY**

The primary objective of this investigation was to establish the perception of the agile methodology in South Africa. The research was done through a web-based survey questionnaires and the participants were invited via email. A random sample of the respondents was selected for an informal interviews.

A total of 905 invitations were sent and 181 responses were received giving the response rate of 20%. The response rate for the different groups were as followed; software developers 21.0%, managers 11.6% software engineers 14.4%, testers 14.9%, architects 12.7%, analyst 17.7%, infrastructure and software support 7.7%.

The study was conducted over a period of four weeks from April 8, 2013 to May 6, 2013 and the respondents were asked 30 questions. The questions were divided into two sections; personal section that dealt with qualification, role in the team and development experience, agile development section that dealt with perception, successes and failures of the agile methodology.

In the personal section, the questions were designed to understand who the respondents are and their qualification, role in the team and professional experience. In the agile development section, the questions were designed to better understand the respondents' perceptions of the agile methodology. The respondents were asked if they like the agile methodology, whether they understand the concept behind the agile methodology, whether the methodology is working well or poorly in their teams, what are the successes and failures of the methodology. In this section, the respondents had the freedom to express their personal opinion.

### **4. RESULTS AND DISCUSSION**

From the results, it was established that the number of respondents who have or had an excellent or good and the number of respondents who have or had a poor or very poor experience with the agile methodology were equal. The trend also indicated that respondents with degree tend to favour the adoption of the agile methodology while respondents without degree tend to disfavour the adoption of the agile methodology.

Table 1. Respondents per group

Role	No of respondents	Percentage
Software developers	38	21.0
Managers	21	11.6
Software engineers	26	14.4
Testers	27	14.9
Architects	23	12.7
Analysts	32	17.7
Infrastructure & support	14	7.7

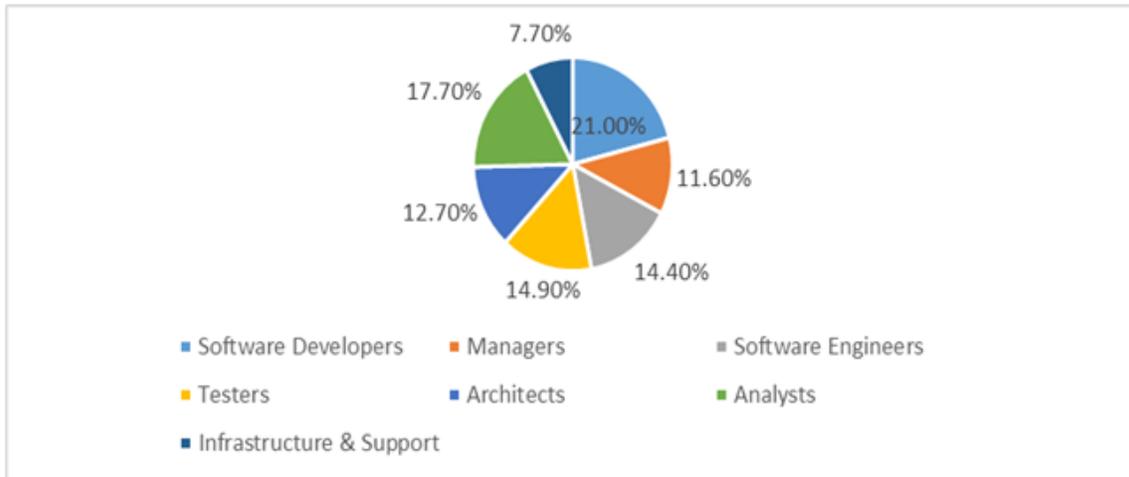


Figure 1. Respondents per group

Table 2. Qualification per respondents

Role	No degree	Undergraduate degree	Postgraduate degree
Software developers	22	12	4
Managers	10	10	1
Software engineers	6	12	8
Testers	18	9	0
Architects	11	10	2
Analysts	11	20	1
Infrastructure & support	13	1	0

Table 3. Work experience per respondents

Duration	No of respondents	Percentage
0 – 6 months	26	14.4
6 months – 2 years	50	27.6
2 years – 5 years	57	31.5
5 years +	48	26.5

Table 4. Usage of agile methodology

Usage	No of respondents	Percentage
Currently using	108	59.7
Once used	43	23.8
Never used	30	16.5

Table 5. Software developers' experience with agile

Experience	No of respondents	Percentage
Excellent	2	5.3
Good	14	36.8
Poor	10	26.3
Very poor	4	10.5
Don't know	8	21.1

Table 6. Managers' experience with agile

Experience	No of respondents	Percentage
Excellent	0	0
Good	8	38.1
Poor	8	38.1
Very poor	5	23.8
Don't know	0	0

Table 7. Software engineers' experience with agile

Experience	No of respondents	Percentage
Excellent	8	30.8
Good	10	38.5
Poor	6	23.1
Very poor	2	7.6
Don't know	0	0

Table 8. Testers' experience with agile

Experience	No of respondents	Percentage
Excellent	0	0
Good	9	33.3
Poor	9	33.3
Very poor	6	22.3
Don't know	3	11.1

Table 9. Architects' experience with agile

Experience	No of respondents	Percentage
Excellent	1	4.4
Good	8	34.8
Poor	8	34.8
Very poor	1	4.4
Don't know	5	21.6

Table 10. Analysts' experience with agile

Experience	No of respondents	Percentage
Excellent	4	12.5
Good	16	50.0
Poor	6	18.8
Very poor	4	12.5
Don't know	2	6.2

Table 11. Infrastructure & support experience with agile

Experience	No of respondents	Percentage
Excellent	0	0
Good	0	0
Poor	3	21.4
Very poor	3	21.4
Don't know	8	57.2

Table 12. General experience with agile methodology

Experience	No of respondents	Percentage
Excellent	15	8.3
Good	65	35.9
Poor	50	27.6
Very poor	25	13.8
Don't know	26	14.4

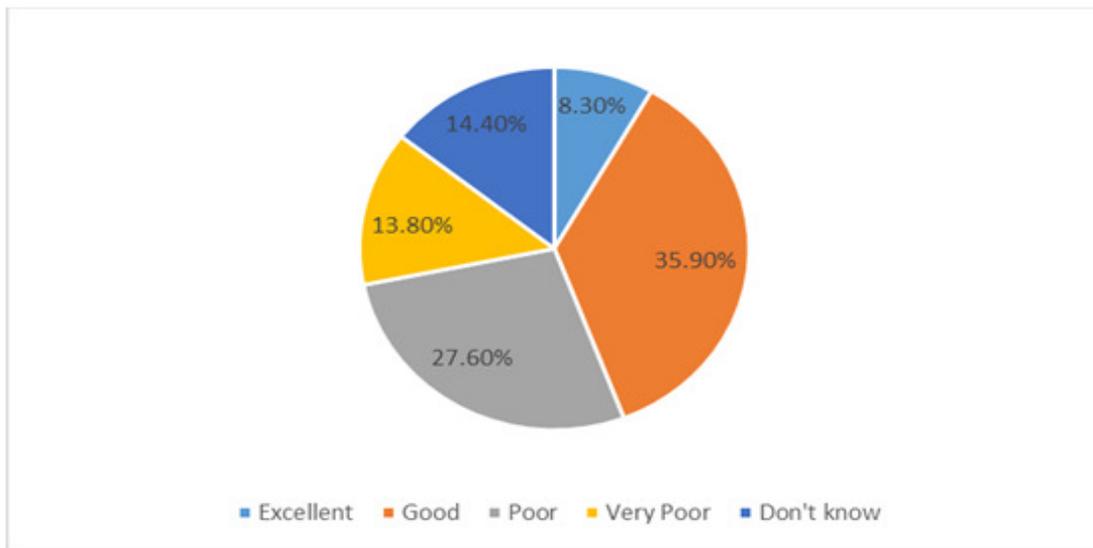


Figure 2: General experience with agile methodology

Table 13. Software developers' interest in agile

<b>Interested</b>	<b>No of respondents</b>	<b>Percentage</b>
Very interested	3	7.9
Interested	9	23.7
Not interested	18	47.4
Don't know	8	21.0

Table 14. Managers' interest in agile

<b>Interested</b>	<b>No of respondents</b>	<b>Percentage</b>
Very interested	2	9.5
Interested	8	38.1
Not interested	6	28.6
Don't know	5	23.8

Table 15. Software engineers' interest in agile

<b>Interested</b>	<b>No of respondents</b>	<b>Percentage</b>
Very interested	7	26.9
Interested	16	61.5
Not interested	2	7.8
Don't know	1	3.8

Table 16. Testers' interest in agile

<b>Interested</b>	<b>No of respondents</b>	<b>Percentage</b>
Very interested	0	0
Interested	5	18.5
Not interested	14	51.9
Don't know	8	29.6

Table 17. Architects' interest in agile

<b>Interested</b>	<b>No of respondents</b>	<b>Percentage</b>
Very interested	2	8.7
Interested	5	21.7
Not interested	13	56.6
Don't know	3	13.0

Table 18. Analysts' interest in agile

<b>Interested</b>	<b>No of respondents</b>	<b>Percentage</b>
Very interested	5	15.6
Interested	13	40.6
Not interested	11	34.4
Don't know	3	9.4

Table 19: Infrastructure &amp; support's interest in agile

Interested	No of respondents	Percentage
Very interested	0	0
Interested	0	0
Not interested	10	71.4
Don't know	4	28.6

Table 20: General interest in agile methodology

Interested	No of respondents	Percentage
Very interested	19	10.5
Interested	56	30.9
Not interested	74	40.9
Don't know	32	17.7

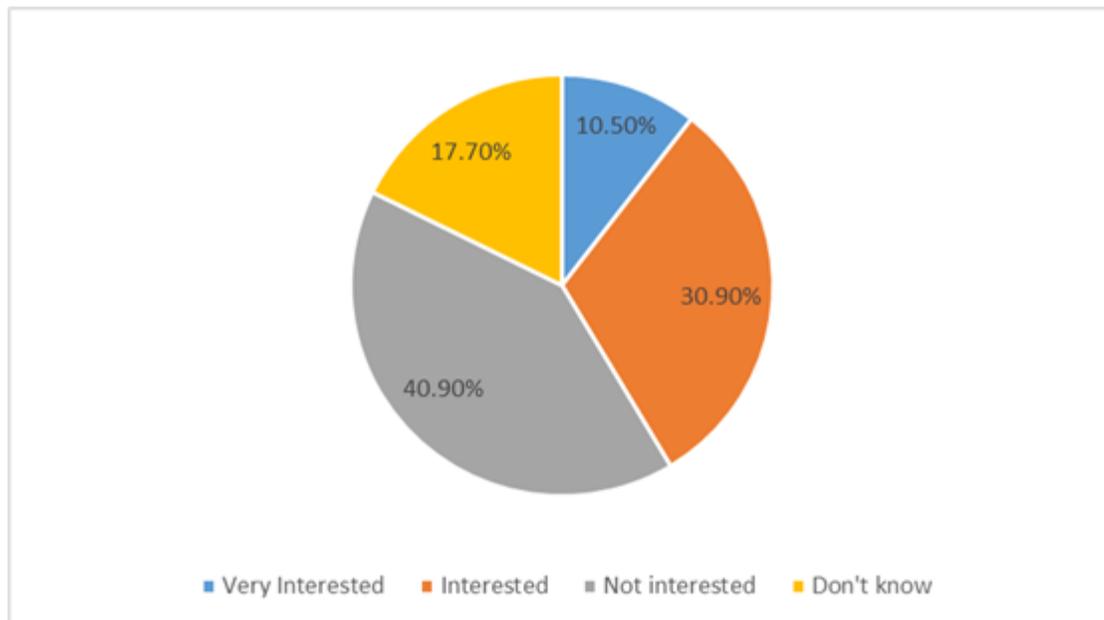


Figure 3: General interest in agile methodology

## 5. CONCLUSION AND FUTURE WORK

The main objective of this research was to investigate the perception of the agile methodology in the South African software development industry. The main findings were that 8.3% of the respondents have or had an excellent experience with the adoption of the agile methodology in their teams and 35.9% have or had a good experience while 27.6% have or had a poor experience with only 13.8% having a very poor experience with the adoption of the agile methodology.

Looking at the interest in adopting the methodology in the teams, 10.5% were very interested, 30.9% were interested while 40.9% were not interested and 17.7 were undecided.

Analysis of the results showed that software engineers are more in favour of adopting the agile methodology with 61.5% while 71.4% of infrastructure and support respondents do not support the adoption of the methodology in their teams.

The results gives a clear indication of how the agile adoption is perceived in South Africa. The follow up to this work could be an in-depth investigation of the level of adoption and penetration of the agile methodology in the South African software development industry, what are the failures and successes, what are the limitations and benefits.

In general, there was an impression among the respondents especially the developers that, there is not a single way to practice agile methodology and if a particular methodology was to be followed in the development of software, this methodology has to be followed to the latter, if not, there will be unforeseen consequences. This is somehow absurd as any agile methodology is supposed to be adaptive to the project needs and changes [6].

## **ACKNOWLEDGEMENTS**

We would like to thank all those respondents who took time to respond to the questions and their feedback that made this paper possible. We would also like to give special thanks to Prof. Barry Dwolatzky from the University of the Witwatersrand for his contribution in guiding us into how to conduct research when we were master students at the university.

## **REFERENCES**

- [1] A. Begel, N. Nagappan. Usage and Perceptions of Agile Software Development in an Industrial Context: An Exploratory Study, ESEM '07 Proceedings of the First International Symposium on Empirical Software Engineering and Measurements, 2007.
- [2] Agile Manifesto. Twelve Principles of Agile Software, <http://agilemanifesto.org>. Last accessed on April 4, 2013.
- [3] Alex Colburn Jonathan Hsieh Matthew Kehrt Aaron Kimball. There is no software engineering crisis. January 16, 2008.
- [4] S.Nithila, K. Priyadharshani, Y. S. G. Attanayake, T. Arani and C.D. Manawadu. Emergence of Agile Methodologies: "Perceptions from Software Practitioners in Sri Lanka". International Journal of Scientific and Research Publications, Volume 3, Issue 11, November 2013.
- [5] A. Asnawi, A. Gravell, G. Wills. An Empirical Study: Understanding Factors and Barriers for implementing Agile Methods in Malaysia, IDoESE'10, Sept. 2010.
- [6] C. Tsun and C. Dac-Buu, "A Survey Study of Critical Success Factors in Agile Software Projects", Journal of Systems and Software, vol. 81, pp. 961-971, June 2008.
- [7] A. Asnawi, A. Gravell, G. Wills. An Empirical Study: Understanding Factors and Barriers for implementing Agile Methods in Malaysia, IDoESE'10, Sept. 2010.

*INTENTIONAL BLANK*

# ADVANCED CLOUD PRIVACY THREAT MODELING

Ali Gholami and Erwin Laure

HPCViz Department,  
KTH Royal Institute of Technology, Stockholm, Sweden  
{gholami, erwinl@pdc.kth.se}

## ABSTRACT

*Privacy-preservation for sensitive data has become a challenging issue in cloud computing. Threat modeling as a part of requirements engineering in secure software development provides a structured approach for identifying attacks and proposing countermeasures against the exploitation of vulnerabilities in a system. This paper describes an extension of Cloud Privacy Threat Modeling (CPTM) methodology for privacy threat modeling in relation to processing sensitive data in cloud computing environments. It describes the modeling methodology that involved applying Method Engineering to specify characteristics of a cloud privacy threat modeling methodology, different steps in the proposed methodology and corresponding products. We believe that the extended methodology facilitates the application of a privacy-preserving cloud software development approach from requirements engineering to design.*

## KEYWORDS

*Threat Modeling, Privacy, Method Engineering, Cloud Software Development*

## 1. INTRODUCTION

Many organizations that handle sensitive information are considering using cloud computing as it provides easily scalable resources and significant economic benefits in the form of reduced operational costs. However, it can be complicated to correctly identify the relevant privacy requirements for processing sensitive data in cloud computing environments due to the range of privacy legislation and regulations that exist. Some examples of such legislation are the EU Data Protection Directive (DPD) [1] and the US Health Insurance Portability and Accountability Act (HIPAA) [2], both of which demand privacy-preservation for handling personally identifiable information.

Threat modeling is an important part of the process of developing secure software – it provides a structured approach that can be used to identify attacks and to propose countermeasures to prevent vulnerabilities in a system from being exploited [3]. However, the issues of privacy and security are really two distinct topics [4] as security is a core privacy concept, and the current focus of the existing threat modeling methodologies is not on privacy in cloud computing, which makes it difficult to apply these methodologies to developing privacy-preserving software in the context of cloud computing environments.

In 2013, the Cloud Privacy Threat Modeling (CPTM) [6] methodology was proposed as a new threat modeling methodology for cloud computing. The CPTM approach was originally designed to support only the EU DPD, for reducing the complexity of privacy threat modeling. Additionally, there were weaknesses in threat identification step through architectural designs in the early stages of Software Development Life Cycle (SDLC) that demanded improvements.

This paper describes an extension of the CPTM methodology according to the principles of Method Engineering (ME) [5]. The method that has been applied is one known as “Extension-based”, which is used for enhancing the process of identifying privacy threats by applying meta-models/patterns and predefined requirements. This new methodology that is being proposed provides strong methodological support for privacy legislation and regulation in cloud computing environments. We describe the high-level requirements for an ideal privacy threat modeling methodology in cloud computing, and construct an extension of CPTM by applying the requirements that were identified.

The rest of this paper is organized as follows. Section 2 provides a background to these developments by outlining the CPTM methodology and existing related work. Section 3 describes the characteristics that are desirable in privacy threat modeling for cloud computing environments. Section 4 describes the steps and products for the proposed new methodology. Section 5 presents the conclusions from this research and directions for future research.

## **2. BACKGROUND AND RELATED WORK**

The CPTM [6] methodology was proposed as a specific privacy-preservation threat modeling methodology for cloud computing environments that process sensitive data within the EU’s jurisdiction. The key differences between the CPTM methodology and other existing threat modeling methodologies are that CPTM provides a lightweight methodology as it encompasses definitions of the relevant DPD [1] requirements, and in addition that it incorporates classification of important privacy threats, and provides countermeasures for any threats that are identified.

For the first step in the CPTM approach, the DPD terminology is used to identify the main entities to cloud environments that are in the process of being developed. Secondly, the CPTM methodology describes the privacy requirements that must be implemented in the environment, e.g., lawfulness, informed consent, purpose binding, data minimization, data accuracy, transparency, data security, and accountability. Finally, the CPTM approach provides countermeasures for the identified threats. Detailed description of the CPTM methodology steps have been discussed in [6] (Sections 3, 4 and 5).

While the CPTM methodology was the first initiative for privacy threat modeling for cloud computing environments in accordance with the EU’s DPD, it nevertheless does not support other privacy legislation, such as that required under the HIPAA [2]. In this paper, we identify the CPTM methodology weaknesses such as support for different privacy legislation and threat identification process and refine the methodology by applying an Extension-based ME approach.

There has been a significant amount of research in the area of threat modeling for various information systems with the goal of identifying a set of generic security threats [7], [8], and [9]. There are guidelines for reducing the security risks associated with cloud services, but none of

these include an outline of privacy threat modeling. The Cloud Security Alliance (CSA) guidelines [10] are not thorough enough to be referred as a privacy threat model because they are not specific to privacy-preservation.

The European Network and Information Security Agency (ENISA) has identified a broad range of both security risks and benefits associated with cloud computing, including the protection of sensitive data [11]. Pearson [4] describes the key privacy challenges in cloud computing that arise from a lack of user control, a lack of training and expertise, unauthorized secondary usage, complexity of regulatory compliance, trans-border data flow restrictions, and litigation.

LINDDUN [12] is an approach to privacy modeling that is short for “likability, identifiability, non-repudiation, detectability, information disclosure, content unawareness, and non-compliance”. This approach proposes a comprehensive generic methodology for the elicitation of privacy requirement through mapping initial data flow diagrams of application scenarios to the corresponding threats. The Commission on Information Technology and Liberties (CNIL) has proposed a methodology for privacy risk management [13] that may be used by information systems that must comply with the DPD.

### **3. CHARACTERISTICS OF A PRIVACY THREAT MODELING METHODOLOGY FOR CLOUD COMPUTING**

This section describes the features that we believe a privacy threat model should have in order to be used for developing privacy-preserving software in clouds in an efficient manner. Based on the properties that are identified, we then apply the Extension-based methodology design approach to construct an extension of the CPTM for supporting various privacy legislation in Section 4.

#### **3.1. Privacy Legislation Support**

Methodological support for the regulatory frameworks that define privacy requirements for processing personal or sensitive data is a key concern. Privacy legislation and regulations can become complicated for cloud customers and software engineering teams, particularly because of the different terminologies in use in the IT and legal fields. In addition, privacy threat modeling methodologies are not emphasized in existing threat modeling methodologies, which causes ambiguity for privacy threat identification.

#### **3.2. Technical Deployment and Service Models**

Cloud computing delivers computing software, platforms and infrastructures as services based on pay-as-you-go models. Cloud service models can be deployed for on-demand storage and computing power can be provided in the form of software-as-a-service (SaaS), platform-as-a-service (PaaS) or infrastructure-as-a-service (IaaS) [14]. Cloud services can be delivered to consumers using different cloud deployment models: private cloud, community cloud, public cloud, and hybrid cloud. Table 1, outlines the five essential characteristics of cloud computing [14].

### 3.3. Customer Needs

The actual needs of the cloud consumers must be taken into consideration throughout the whole life cycle of a project. Additionally, during the course of a project, requests for changes often arise and these may affect the design of the final system. Consequently it is important to identify any privacy threats arising from the customer needs that result from such change requests. Customer satisfaction can be achieved through engaging customers from the early stages of threat modeling so that the resulting system satisfies the customer's needs while maintaining adequate levels of privacy.

### 3.4. Usability

Cloud-based tools aim at reducing IT costs and supporting faster release cycles of high quality software. Threat modeling mechanisms for cloud environments should therefore be compatible with the typical fast pace of software development in clouds-based projects. However producing easy-to-use products with an appropriate balance between maintaining the required levels of privacy while satisfying the consumer's demands can be challenging when it comes to cloud environments.

### 3.5. Traceability

Each potential threat that is identified should be documented accurately and be traceable in conjunction with the associated privacy requirements. If threats can be traced in this manner, it means that threat modeling activities are efficient in tracing of the original privacy requirements that are included in the contextual information and changes over the post-requirement steps such as design, implementation, verification and validation.

Table 1, The five essential characteristics of cloud computing [14]

Cloud Characteristic	Description	Application
<b>On-demand self-service</b>	For automatically providing a consumer with provisioning capabilities as needed.	Server, Time, Network and Storage
<b>Broad network access</b>	For heterogeneous thin or thick client platforms.	Smartphones, tablets, PCs, wide range of locations
<b>Resource pooling</b>	The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model.	Physical and virtual resources with dynamic provisioning
<b>Rapid elasticity</b>	Capabilities can be elastically provisioned and released, in some cases automatically, to scale rapidly outward and inward with demand.	Adding or removing nodes, servers, resource or instances

<b>Measured service</b>	Automated control and optimization of a resource through measuring or monitoring services for various reasons, including billing, effective use of resources, or predictive planning.	Storage, processing, billing, , bandwidth, and active user accounts
-------------------------	---	---

#### 4. METHODOLOGY STEPS AND THEIR PRODUCTS

Motivated by the facts that privacy and security are two distinct topics and that no single methodology could fit all possible software development activities, we apply ME that aims to construct methodologies to satisfy the demands of specific organizations or projects [17]. In [5], ME is defined as “the engineering discipline to design, construct, and adapt methods, techniques and tools for the development of information systems”.

There are several approaches to ME [17], [15] such as a fundamentally “ad-hoc” approach where a new method is constructed from scratch, “paradigm-based” approaches where an existing meta-model is instantiated, abstracted or adapted to achieve the target methodology, “Extension-based” approaches that aim to enhance an existing methodology with new concepts and features, and “assembly-based” approaches where a methodology is constructed by assembling method fragments within a repository.

Figure 1 represents different phases in a common SDLC. Initial security requirements are collected and managed in the requirements engineering phase (A). This includes identifying the quality attributes of the project and assessing the risk associated with achieving them. A design is composed of architectural solution, attack surface analysis and the privacy threat model. Potential privacy threats against the software that is being developed are identified and solutions are proposed to mitigate for adversarial attacks (B). The proposed solution from the design phase is implemented through a technical solution and deployment (C). This includes performing static analysis on source code for software comprehension without actually executing programs. The verification process (D) includes extensive testing, dynamic analysis on the executing programs on virtual resources and fuzzing as a black-box testing approach to discover coding errors and security loopholes in the cloud system. Finally, in the Validation phase the end-users participate to assess the actual results versus their expectations, and may put forth further change requests if needed.

Our proposed methodology identifies the privacy requirements in the Requirements Engineering step, as shown in Figure 2. The results from the Requirements Engineering, which include specifications for privacy regulatory compliance, are fed into the Design step, where activities such as specifying the appropriate cloud environment, identifying privacy threats, evaluating risks and mitigating threats are conducted. Then the produced privacy threat model would be used in the implementation step finally it would be verified and validated in the subsequent steps.

Cloud stakeholders and participants such as cloud users, software engineering team and legal experts will engage in the activities shown in Figure 2 to implement the threat model in context of steps A and B in Figure 1. Cloud software architect as a member of the software engineering team initiates a learning session to clarify the methodology steps and their products, privacy requirements (introducing the law title that is needed to be enforced in the cloud environment), and quality attributes such as performance, usability. The legal experts will identify the definitive

requirements that ensure the privacy of data in the platform. In the Design step, the cloud software architect presents architecture of the developing cloud environment for various participants. This will result in a unified terminology to be used in the privacy threat model.

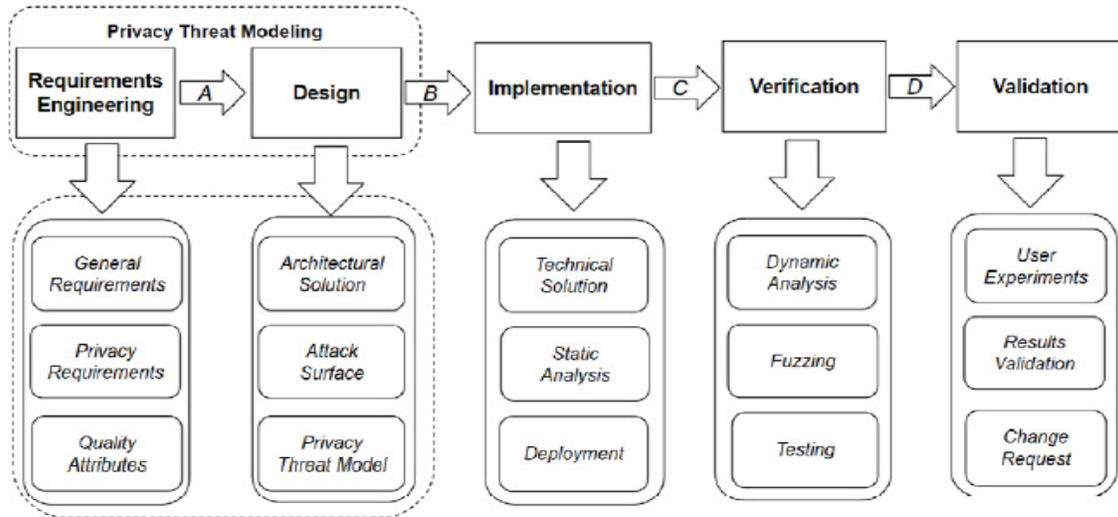


Figure 1, Privacy Threat Modeling in Requirements Engineering and Design of a SDLC

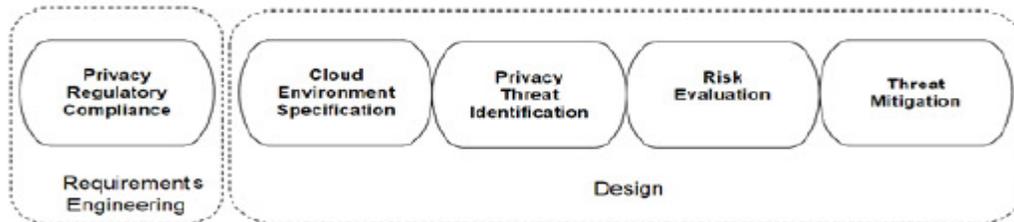


Figure 2, Overview of the Extended CPTM Methodology Steps

The rest of this section outlines the implementation model of the steps represented in Figure 2.

### 3.1 Privacy Regulatory Compliance

Interpreting privacy regulatory frameworks can be often complex for software engineering teams. In the privacy regulatory compliance step, learning sessions with privacy experts, end-users and requirements engineers facilitates the elicitation of privacy requirements (PR). For example, in the EU DPD some of the privacy requirements are: lawfulness, informed consent, purpose binding, transparency, data minimization, data accuracy, data security, and accountability [6]. Each of the requirements that are identified will be labeled with an identifier, e.g., (PRi), name and description to be used in later stages.

### 3.2 Cloud Environment Specification

To ensure that the final cloud software will comply with the relevant legal and regulatory framework, several of the key characteristics that are affected by cloud computing services (including virtualization, outsourcing, offshoring, and autonomic technologies) must be specified.

For this purpose, the physical/logical architectures of the deployment and service model can be developed according to the following steps

- **Step A:** Define the cloud actors [18] (such as *Cloud Consumer, Cloud Provider, Cloud Auditor, Cloud Broker, and Cloud Carrier*). Cloud consumer is a person or organization uses service from cloud providers in context of a business relationship. Cloud provider makes service available to interested users. Cloud auditor conducts independent assessment of cloud services, operations, performance and security of the deployment. Cloud broker manages the use, performance and delivery of cloud services and establishes relationships between cloud providers and cloud consumers. Cloud carrier provides connectivity and transport of cloud services from cloud providers to cloud consumers through the network.
- **Step B:** Describe a detailed model of the cloud deployment physical architecture where the components will be deployed across the cloud infrastructure. This should give details of where the components will be deployed and run, for example, the operating system version, the database version, the virtual machine location, and where the database server will run.
- **Step C:** Describe the logical architecture of the cloud services model where the major cloud services, along with and the relationships between them that are necessary to fulfill the project requirements, are recorded. This should include the data flow and connections between the relevant cloud services and actors. Note that in this context, an entity is a cloud service with a set of properties that meet a specific functional requirement.
- **Step D:** Describe the assets that need to be protected, the boundaries of the cloud and any potential attackers that might endanger either the cloud environment or the assets that have been identified as being associated with that particular cloud.

The cloud environment specification step consists of composing an architectural report including assets that are subject to privacy protection, cloud actors, physical architecture of the deployment model, and logical architecture of the service model.

### 3.3 Privacy Threat Identification

In this step, privacy threats against the PRs that were established in section 3.1 will be identified and analyzed. To achieve this, the system designers will undertake the following steps.

- **Step A:** Select a privacy requirement from the PR list for threat analysis, e.g., (PR2).
- **Step B:** Correlate identified cloud actors (Step A from Section 3.2) with the actor roles that are defined in the project's privacy law. For example, correlating *the Data Controller role as a Cloud Consumer, or the Data Processor role as a Cloud Provider* in the DPD.
- **Step C:** Identify all the technical threats that can be launched by an adversary to privacy and label them in the specified cloud environment. Each identified threat can be named as a  $T(i,j)$ , where  $i$  indicates that threat  $T$  that corresponds to  $PR_i$  and  $j$  indicates the

actual threat number. For example, in T(2,5) 2 indicates relevance of the threat to PR2 and 5 is the actual threat number.

- **Step D:** Repeat the previous steps until all PRs are processed.

The threat identification step consists of composing an analysis report including a list of threats including id, name, date, author, threat scenario for each class of the PRs.

### 3.4 Risk Evaluation

In this step, all actors participate to rank the threats that have been identified in Section 3.3 with regard to their estimated level of importance and the expected severity of their effect on the overall privacy of the cloud environment. The Importance indicates the likelihood of a particular threat occurring and the level of the Effect indicates the likely severity of the damage if that threat against the cloud environment were carried out.

Assume there are three identified PRs (PR<sub>1</sub>, PR<sub>2</sub>, PR<sub>3</sub>) in addition to related privacy threats T(1,4), T(2,1) and T(3,3) from previous steps for an imaginary cloud system. In this imagined cloud, various participants in the project such as Alice (Cloud Consumer), Bob (Cloud Provider), Dennis (Software Architect), Tom (Lawyer) and Rosa (Cloud Carrier) evaluate the corresponding risk of each identified threats, as illustrated in Table 2.

Table 2, Prioritization of the identified threats, L (Low), M (Moderate), H(High)

ID	Name	Scenario	Importance	Effect	Participants
T(1,4)	Data Accumulation over Time	The cloud system stores a huge amount of data from Cloud Consumers over the time. This can be done through extensive analysis over collected data from different sources.	H	M	Alice, Bob, Dennis, Tom
T(2,1)	Linkability of Records	A record owner can be linked through the adversarial background knowledge for the published data to the Cloud Provider.	H	H	Alice, Tom, Bob
T(3,3)	Cross-linking of data processing	A Cloud Consumer is able to run cross-linking queries over multiple data sets from different data sources.	M	H	Tom, Bob, Rosa

This step results in composing a risk evaluation report similar to the example in Table 2. This report prioritizes the importance and effects of the privacy threats and it will be used in the Threat Mitigation step in Section 3.5.

### **3.5 Threat Mitigation**

In this step, the threat modeling team propose countermeasures to the threats that were identified in the previous step as having the highest likelihood of occurrence and the worst potential effects on the cloud environment. Each countermeasure should clearly describe a solution that reduces the probability of the threat occurring and that also reduces the negative effects on the cloud if the threat was carried out.

Finally, the recommended countermeasures from this step should be documented and fed into the implementation step to be realized through coding and for their effectiveness to be assessed by static analysis. In the later stages of verification and validation, each such countermeasure will be evaluated and approved by the participants.

## **5. CONCLUSIONS AND FUTURE WORK**

In this paper we identified the requisite steps to build a privacy threat modeling methodology for cloud computing environments using an Extension-based Method Engineering approach. For this purpose, we extended the Cloud Privacy Threat Modeling (CPTM) methodology to incorporate compliance with various legal and regulatory frameworks, in addition to improving the threat identification process.

In future research, we aim to apply the proposed methodology within domain independent clouds that process sensitive data. This will validate our methodology for providing customized privacy threat modeling for other privacy regulations, such as HIPAA, in cloud computing environments.

### **ACKNOWLEDGEMENTS**

This work funded by the EU FP7 project Scalable, Secure Storage and Analysis of Biobank Data under Grant Agreement no. 317871.

### **REFERENCES**

- [1] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal (OJ) 1995, L 281, p. 31.
- [2] Centers for Medicare and Medicaid Services, “The Health Insurance Portability and Accountability Act of 1996 (HIPAA)”, 1996.
- [3] F. Swiderski and W. Snyder, Threat Modeling, Microsoft Press, 2003.
- [4] S. Pearson, “Privacy, Security and Trust in Cloud Computing”, Computer Communications and Networks, Springer London, pp 3-42, 2013.

- [5] S. Brinkkemper, "Method engineering: engineering of information systems development methods and tools", *Information and Software Technology*, Vol. 38, No. 4, 1996, pp. 275-280
- [6] A. Gholami, A.-S. Lind, J. Reichel, J.-E. Litton, A. Edlund, and E. Laure, "Privacy threat modeling for emerging biobankclouds," *Procedia Computer Science*, vol. 37, pp. 489-496, 2014.
- [7] B. Schneier, "Threat Modeling and Risk Assessment", *View* (2000), 214-229.
- [8] Y. Chen, "Stakeholder Value Driven Threat Modeling for Off the Shelf Based Systems", *IEEE Computer Society* 2007, 91-92.
- [9] S. Baek, J. Han, Y. Song, "Security Threat Modeling and Requirement Analysis Method Based on Goal-Scenario", *Springer Netherlands* 2012, 419-423.
- [10] The Cloud Security Alliance (CSA). Security guidance for critical areas of focus in cloud computing v3.0, (2011), <https://cloudsecurityalliance.org/guidance/csaguide.v3.0.pdf>, visited October 2013.
- [11] D. Catteddu and G. Hogben, "Cloud computing. Benefits, risks and recommendations for information security", *ENISA Report*, 2009.
- [12] M. Deng, W. Kim, R. Scandariato, B. Preneel and W. Joosen, "A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements", *Requir. Eng.*, 2011, 3-32.
- [13] CNIL. Methodology for Privacy Risk Management, (2012). Available at: <http://www.cnil.fr/fileadmin/documents/en/CNILManagingPrivacyRisksMethodology.pdf>, visited October 2013.
- [14] NIST SP 800-145, "A NIST definition of cloud computing", [online] <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.
- [15] R. Rahimian and R. Ramsin, "Designing an agile methodology from mobile software development: a hybrid method engineering approach," in *2nd Int. Conf. on Research Challenges in Information Science*, Marrakech, pp. 337- 342, 2008.
- [16] K. Kumar, R. J. Welke, "Method Engineering: a proposal for situation-specific methodology construction", in *Systems Analysis and Design: A Research Agenda*, 1992.
- [17] J. Ralyté, R. Deneckère, C. Rolland, "Towards a generic model for situational method engineering", in *Proc. of CAiSE'03 (LNCS 2681)*, 2003, pp. 95-110.
- [18] R. B. Bohn, J. Messina, F. Liu, J. Tong, and J. Mao, "NIST cloud computing reference architecture," in *2011 IEEE World Congress on Services*. IEEE Computer Society, 2011, pp. 594-596.

## AUTHORS

**Ali Gholami** is a PhD student at the KTH Royal Institute of Technology. His research interests include the use of data structures and algorithms to build adaptive data management systems. Another area of his research focuses on the security concerns associated with cloud computing. He is currently exploring strong and usable security factors to enable researchers to process sensitive data in the cloud.



**Professor Erwin Laure** is Director of the PDC - Center for High Performance Computing Center at KTH, Stockholm. He is the Coordinator of the EC-funded "EPiGRAM" and "ExaFLOW" projects as well as of the HPC Centre of Excellence for Bio-molecular Research "BioExcel" and actively involved in major e-infrastructure projects (EGI, PRACE, EUDAT) as well as exascale computing projects. His research interests include programming environments, languages, compilers and runtime systems for parallel and distributed computing, with a focus on exascale computing.



*INTENTIONAL BLANK*

# A TAXONOMY FOR TOOLS, PROCESSES AND LANGUAGES IN AUTOMOTIVE SOFTWARE ENGINEERING

Florian Bock<sup>1</sup> and Daniel Homm<sup>1</sup> and Sebastian Siegl<sup>2</sup> and  
Reinhard German<sup>1</sup>

<sup>1</sup>Department of Computer Science 7,  
Friedrich-Alexander-University, 91058 Erlangen, Germany  
florian.inifau.bock@fau.de,daniel.homm@fau.de,reinhard.german@fau.de

<sup>2</sup>Audi AG,  
85045 Ingolstadt, Germany  
sebastian.siegl@audi.de

## **ABSTRACT**

*Within the growing domain of software engineering in the automotive sector, the number of used tools, processes, methods and languages has increased distinctly in the past years. To be able to choose proper methods for particular development use cases, factors like the intended use, key-features and possible limitations have to be evaluated. This requires a taxonomy that aids the decision making. An analysis of the main existing taxonomies revealed two major deficiencies: the lack of the automotive focus and the limitation to particular engineering method types. To face this, a graphical taxonomy is proposed based on two well-established engineering approaches and enriched with additional classification information. It provides a self-evident and -explanatory overview and comparison technique for engineering methods in the automotive domain. The taxonomy is applied to common automotive engineering methods. The resulting diagram classifies each method and enables the reader to select appropriate solutions for given project requirements.*

## **KEYWORDS**

*Software Engineering, Processes & Tools & Languages, Comparison, Taxonomy, Classification*

## **1. INTRODUCTION**

Since the first definition of the term *Software Engineering* in a NATO conference report from 1968 [29], a lot of new tools, processes, programming languages and other software engineering methods have appeared. They provide different key-features, advantages and disadvantages and they especially differ in their associated application domain. Within these different domains, the automotive sector is the focus of this paper.

Cars have developed from being completely mechanical in the early 20th century to being electromechanical in the subsequent decades until finally reaching the present-day's complexity in terms of hardware and software. Especially in case of software development, such aspects like

the quantity of functions embedded in the car or the binary code size have increased exponentially [10],[11],[13]. To face these challenges, on the one hand, the hardware is continuously improved by more powerful components. On the other hand, the high climax in software challenges cannot be solved just by hardware improvements, but requires evolution in software engineering. The required efforts can be divided into two categories: runtime efforts and design efforts. Runtime efforts are concerned with the optimal execution of complex code on the hardware. Here, software engineering improvements are hardly feasible. Hence, this is not in the focus of this paper. Design efforts relate to the efficient specification of complex software, which results in a need for good software engineering methods. This is the key topic of this paper.

The development cycle for a car series was reduced by about 25% during the past decades [33], while the development complexity increased. Using the same well-established engineering methods would result in a great demand for new man-power, which is not economical. Resources have to be ideally utilized. New software engineering methods can help to reach this goal. However, new methods often differ in several aspects and hence, for each scenario in the development process, different adequate methods are available. To be able to choose the proper approach for a given project scenario, the common methods placed on the market have to be examined, classified and compared to offer this information and classification to potential users. Especially the comparison of methods of fundamentally different types, for example processes and tools, may seem like trying to compare apples and oranges, due to the largely mismatching set of characteristics. Common comparison techniques are not applicable, because they require measurable, quantifiable and matchable characteristics to work properly. Nevertheless, a comparison by any means is necessary to be able to come to a decision for a suitable method in a specific project scenario. Therefore, we introduce a taxonomy, which allows such a classification and is tailored to the automotive domain. We applied it to the main methods available in this area. Thus, a compact and comprehensible overview of the current market situation is also given.

We conducted a survey among 15 representatives from different companies and departments to verify the assumptions established in this paper. It consists of 15 questions. The raw survey data and the survey form can be viewed online [9]. Two-thirds of the respondents work for a car manufacturer, one-fifth in research and the rest for automotive suppliers. Their areas of activity consist of requirements engineering, system architecture, implementation, test, documentation, change-management, administration/organization and miscellaneous topics with an emphasis on requirements engineering and test. The self-evaluation of the respondents regarding their software-engineering skills revealed an overall high average skill level. 47% are decision makers. The age of the respondents ranges from 20 to 49.

## 2. TERMINOLOGY

To be able to describe the classification scheme outlined in this paper, several basic terms have to be taken into account: *Tool*, *Method*, *Process*, *Language* and particularly *General Programming Language* and *Domain Specific Language* [6],[24],[32]. The terms already allow a three-part classification of software engineering approaches into: *Tool* as a piece of software, *Process* as a general description of a procedure, and *Language* as a well-defined mode of communication or specification. The term *Method* is applicable to all of them, because it is a general description of a procedure, which is implied as well in tools, processes and languages.

The classification of available market solutions into this pattern is not always distinct and might require a deep analysis of each approach. There is the possibility that some methods may fit in more than one category.

Terms and subcategories of languages are difficult to determine and apply, because they are partly used quite different depending on the domain or user group. For instance according to [26], languages can be subdivided into *GPLs* and *DSLs*, whereat in [35], *programming-* and *modeling-languages* are employed. As a compromise, the categorization displayed in figure 1 is used below at which *Others* stands for natural languages (e.g. *English*) without any programmatic background.

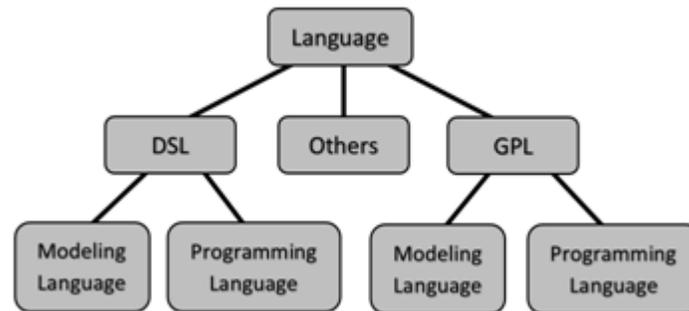


Figure 1. Classification of Languages

The correct classification is not as clear as it might appear at first glance. The main differentiator is obviously the limitation of *DSLs* to a specific domain, whereas *GPLs* can be applied to all domains. Indeed, this is only sufficient as sole distinction feature for some candidates e.g. *C++*, which is clearly a *GPL*. Other languages like the *Unified Modeling Language (UML)* [31] or the *System Modeling Language (SysML)* [30] are apparently limited to a specific domain, but are categorized as a *GPL* [30]. Hence, a more detailed distinction method is required, which can partly be derived from [35] and [26]. This classification task is succinctly described in chapter 4.

### 3. RELATED WORK

There already exist various taxonomies that help to classify software engineering methods. To the best knowledge of the authors, the main approaches have been selected and are elaborated in detail below, with special focus on the applicability to the automotive domain and its requirements.

Blum [8] proposed a classification scheme for engineering methods that distinguishes between *Problem-oriented* and *Product-oriented* attempts as well as between *Conceptual* and *Formal* ones. A matrix of these two differentiation schemes allows a simple classification. However, it does not take into account topics like engineering steps, modeling roles or the automotive context.

Kitchenham [25] focused on the DESMET evaluation method. She identified evaluation types that enable a comparison between different software engineering methods and tools: Quantitative types, qualitative types and other types. The evaluation types are empirical attempts. They need a large amount of data about an engineering method to allow a categorization. This data can only

be obtained for well-defined and ready-to-use methods, which can be tested in real projects. Attempts without any existing application or with limited data are not covered by this taxonomy.

Babar et. al. [5] introduced a taxonomy to compare general software architecture analysis approaches. It consists of 17 evaluation questions grouped in the four categories context, stakeholders, contents and reliability. Their taxonomy is limited to software architecture analysis methods. Additionally, the automotive context is missing.

Hofmeister et. al. [23] proposed a taxonomy for architectural design methods that provides two kinds of comparison techniques: activity-based and artifact-based. The former involves an architectural analysis, synthesis and evaluation, whereas the latter considers architectural concerns or candidate architectural solutions. The taxonomy is lacking the automotive focus.

Broy et. al. [14] defined a taxonomy for engineering tools in the automotive domain. It classifies tools by vertical domain-related and horizontal domain-independent aspects. The former considers language aspects whereas the latter concerns aspects of the tool framework. Prior to the classification of a tool, empirical data has to be obtained by investigating its toolbars/menu items and identifying the underlying functionality as domain-related or -independent. The taxonomy is focused on the automotive domain, however, the limitation to tools excludes languages and processes without an integrated tool.

The main deficiencies of the above summarized approaches are:

- The lack of an automotive focus. Therefore, the results cannot be applied directly to that domain.
- The limitation to a particular type of engineering method. Methods of different types cannot be compared.
- The primarily use of quantifiable characteristics to compare methods. Such approaches are benchmarks with the objective of providing a method ranking. This requires the collection of much data for each method and is only applicable for methods of the same type.

Such limitations are, as already described in the introduction, not feasible in some project settings. Especially at the project start, diverse methods, tools and processes with their individual characteristics are candidates and therefore under investigation. A comparison cannot be accomplished by the above reviewed taxonomies. Hence, a new comparison technique is required, which is developed in this paper as new, generally applicable and lightweight taxonomy for the automotive domain. Its main aim is to guide the decision making by the use of an appropriate overview of the methods in question.

#### **4. TAXONOMY FOR THE AUTOMOTIVE DOMAIN**

Two-thirds of our survey respondents are not satisfied with the methods currently used in their environment. Their willingness to introduce new approaches into their established workflows is quite high. A suitable and lightweight taxonomy that fits to the automotive domain and provides an overview of available methods helps to improve the situation. It may also increase the

willingness of the department to introduce new methods, which is low according to our survey respondents. Such a taxonomy has to be plausible, adaptable to methods of varying types, and clear. It can basically be visualized textually or graphically. If the methods, which should be compared, share the same type and are directly comparable as to e.g. their key features, a textual or tabular approach might be adequate. In the given context, this is obviously not the case. Therefore and by reasons of simplicity and clarity, a graphical taxonomy seems to be the most appropriate way to offer an easy and understandable decision pattern for a wide range of different engineering methods. Primary goal is not to evaluate the performance of the methods and create a ranking, but to offer a lightweight, comprehensible and clear overview and comparison pattern.

As most of the methods commonly used in the automotive domain are based on the *V-Model* [12], it can be taken as a reliable base model. This is also verified by our survey, in which all of the respondents indicate familiarity with it [9]. Though, it is rather generic and therefore neither limited to a specific domain, nor enriched with automotive terms and views. As a result, the automotive context is considered by using a level model that represents the different modelling steps during software development in the automotive domain. Instead of proposing a completely new level model, an already specified and field-tested one is used: the model incorporated in the *EAST-ADL* approach [18] (cf. chapter 5.3). This ensures both adaptability and applicability for the given context.

The level model from the *EAST-ADL*-specification consists of four consecutive abstraction levels [18]:

- *Vehicle Level*: A solution-independent, abstract description of the target car functions (e.g. driver assistance systems). This includes use cases, requirements and high-level descriptions of features- and functions, all of them as graphical as well as textual artifacts.
- *Analysis Level*: A functional black-box decomposition with interface information. The artifacts from the level above are enriched with additional information. The resulting system is designed as a black-box architecture, consisting of several blocks with raw specifications about their interactions, e.g. which information should be collected from outside the system and which output should be returned.
- *Design Level*: A functional white-box decomposition with hardware information, e.g. the type of controller or sensor used in the target system. The black-box specification is filled with the inner behavior in the form of abstract algorithms, state machines and additional information. Thus, a complete system behavior model is created.
- *Implementation Level*: An implementation of the car functions. Here, the system model created in the previous levels is implemented in the target language and delivered to the target platform (for example a controller or another embedded device). The initially defined car functions are practically usable and testable.

Each of the levels contains both specification and test of the particular artifacts.

These levels with their descriptions resemble the phases of the *V-Model*. Hence, the phases and the levels can be overlaid (cf. figure 2). This is valid, because the layered architecture from *EAST-ADL* is derived from the *V-Model* [7].

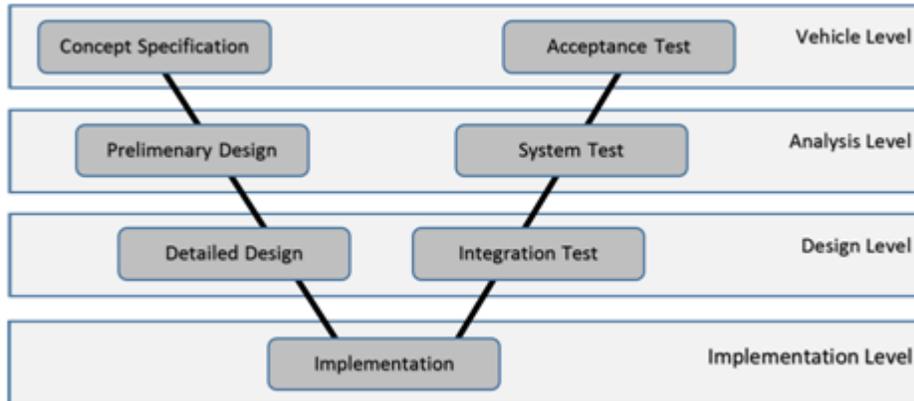
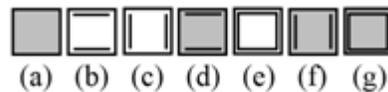


Figure 2. Overlay of the *V-Model* and the *EAST-ADL*-levels

In addition, the type of an engineering method should be reflected in the diagram. As already described in chapter 2, the terms *process*, *tool* and *language* are applicable, whereupon *language* can be subdivided in *DSLs* and *GPLs*. Due to the fact that some methods cover more than one level or step of the *V-Model*, it is not sufficient to simply note a method textually in the diagram. The use of formatted bars as graphical representation for the different methods and their coverage of software development steps seems appropriate.

The lines and the color (in this case gray-scale) of a simple bar are modified in a readable and constructive way to encode the categorization information as combination of *language*, *process* and *tool* (cf. figure 3). This formatting rules ensure that the diagram stays simple and readable.



- |              |                      |                           |
|--------------|----------------------|---------------------------|
| (a) Language | (d) Language/Process | (g) Language/<br>Process/ |
| (b) Process  | (e) Process/Tool     | Tool                      |
| (c) Tool     | (f) Language/Tool    |                           |

Figure 3. Encoding for engineering methods

Additionally, the general type of language should be included in the notation. The background is altered to reflect this information: *dark-gray* for *DSLs* and *light-gray* for *GPLs*. To determine the language type, the classification patterns from [26] and the information from the respective language provider are used.

## 5. EVALUATION

There are several software engineering methods currently available on the market. This paper focuses on the most common and established ones: *Rational Harmony*, *AUTOSAR*, *EAST-ADL*,

*MATLAB/Simulink/TargetLink*, *SCADE*, *ADTF*, *RUP/EUP* and *SimTAny*. We applied our taxonomy to these approaches, which yields their classification depicted in figure 4a/4b. Due to clarity reasons, the phases of the *V-Model* are abbreviated (cf. figure 2) and the approaches are spread across two diagrams.

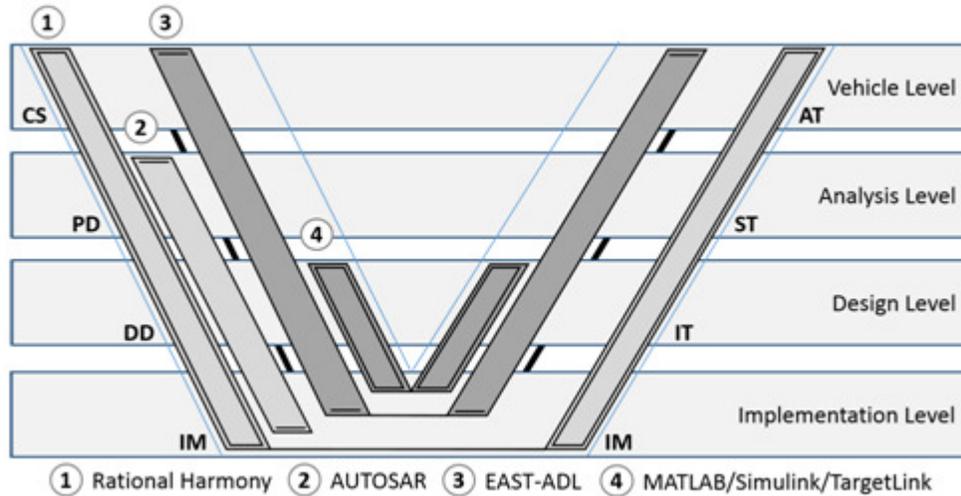


Figure 4a. Automotive specific taxonomy applied to common engineering methods

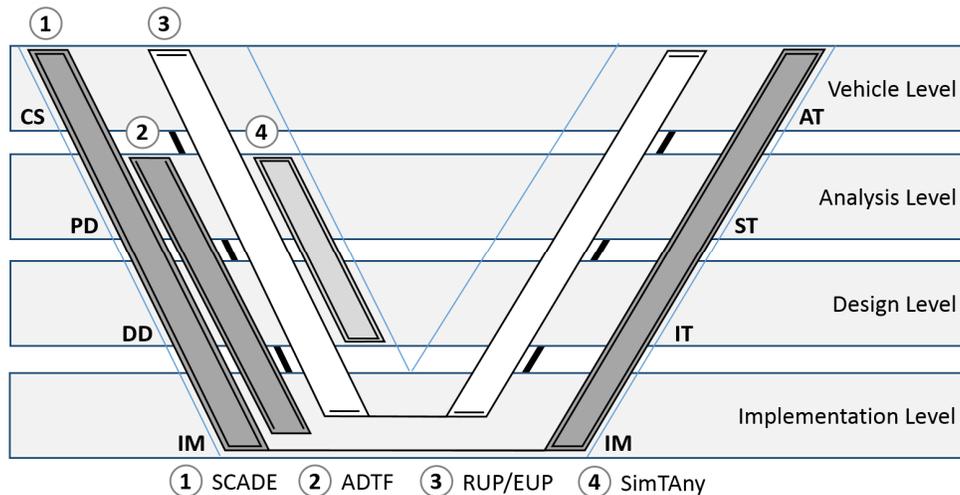


Figure 4b. Automotive specific taxonomy applied to common engineering methods

The diagram can be used to determine an appropriate solution for a given development scenario and to exclude methods, that do not fulfil the project requirements. As depicted in our survey, the knowledge of individual persons and departments about the characteristics of a specific method, its availability or even its existence varies considerably [9]. Our taxonomy deals with this fact by providing an overview with comprehensible information, which can be used without the need for extensive knowledge of each method. This overview also contains the information, whether a tool aspect is included in the method or not. This can be crucial for a reliable decision.

## 5.1. Rational Harmony

*IBM Rational Harmony* [22] is an iterative software modelling process based on the *V-Model* [12]. It is split into two sequenced sections (cf. figure 5). First, the system behavior is modeled as *SysML* model with regard to requirements and use cases. The second step enhances this model and transforms it into an *UML* model, which contains all information necessary to generate both the required system artifacts and the target code. The simulation of the created models and different validation/verification methods are also part of the process and tooling. To increase the usability, semi-automatic wizards assist with the different modelling steps.

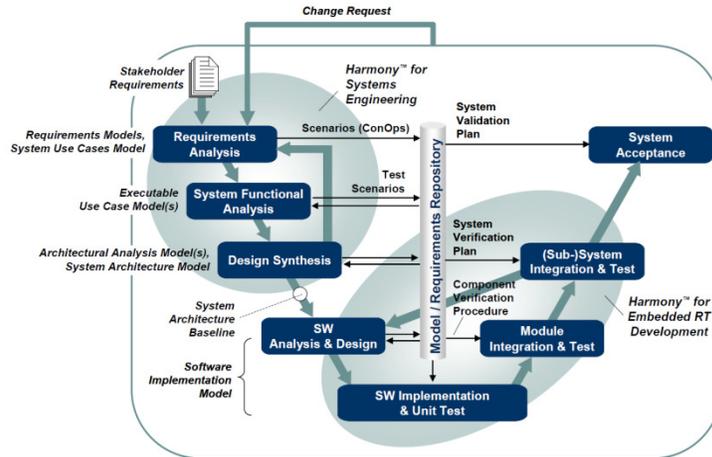


Figure 5. *Harmony* Process Overview [22]

*Rational Harmony* is designed as process with different steps and covers all phases/levels from the taxonomy. As sole implied languages, *UML* and *SysML* are used, which are classified as *GPL*. *Rational Harmony* is always delivered within the tool *Rational Rhapsody*.

Even though being available since 2006 [22], *Rational Harmony* was introduced quite recently in the automotive domain. The implied process steps are generally applicable, so they can easily be adopted for the specific requirements of the domain. Nevertheless, it is not yet widely deployed at present, which is reflected by our survey. Only one-fifth of the respondents indicate the use of *Rational Harmony* in their departments [9].

## 5.2 AUTOSAR

The *AUTomotive Open System ARchitecture* (*AUTOSAR*) [2],[21] is a software architecture standard widely used in the automotive domain and developed by the *AUTOSAR* development partnership. Its focus is the implementation and realization of automotive software systems. To abstract and standardize the development, a layered software architecture is used (cf. figure 6). When utilizing *AUTOSAR*, all required software artifacts for the target car function are located at the *Application Layer*. They consist of so-called *Software Components (SWCs)*, which enfold both the algorithms (which are enclosed in *Runnables*) and the wrapper-code for the car function. To simplify the exchange of model artifacts, a well-defined XML-scheme is used to store all information.

The software architecture models contain abstract as well as low-level information, so the adoption starts within the *Analysis Level* and lasts until the *Implementation Level*. Tests or test strategies are not specified. *AUTOSAR* includes specifications, but no implementation by itself is implied, although external tools exist. Lines of action, which form a process, are provided and *GPL* aspects are available in terms of model definitions. There is, by default, no *DSL* embedded, but an external add-on exists (*ARText* [4]). It is a language framework to build user-defined *DSLs* for *AUTOSAR*.

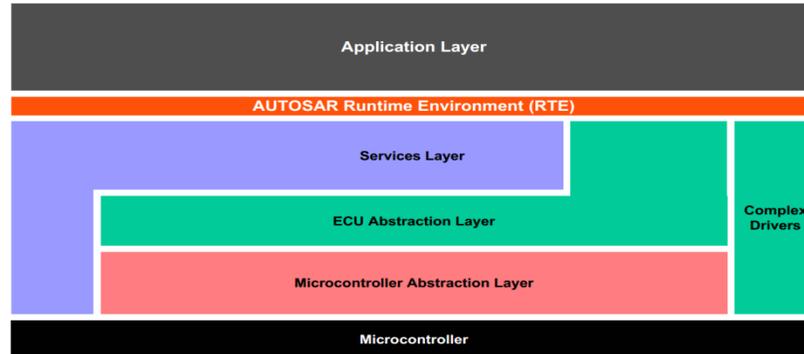


Figure 6. The *AUTOSAR* layered architecture [3]

*AUTOSAR* was initially developed and designed 2005 to be used in the automotive context and is already wide spread in the domain. Our survey shows, that 87% of the respondents are familiar with *AUTOSAR* and 60% already work with it [9].

### 5.3. EAST-ADL

The *Electronics Architecture and Software Technology - Architecture Description Language (EAST-ADL)* [7],[18] is developed and enhanced by the *EAST-ADL Association*. It uses *AUTOSAR* and additionally covers aspects like non-functional requirements, vehicle features and functional/hardware architecture details. The models are categorized in four different abstraction levels (cf. figure 7 and chapter 4). The process starts with a rough initial vehicle model that is enriched during the development, until it is in a highly detailed state and realized as *AUTOSAR* model.

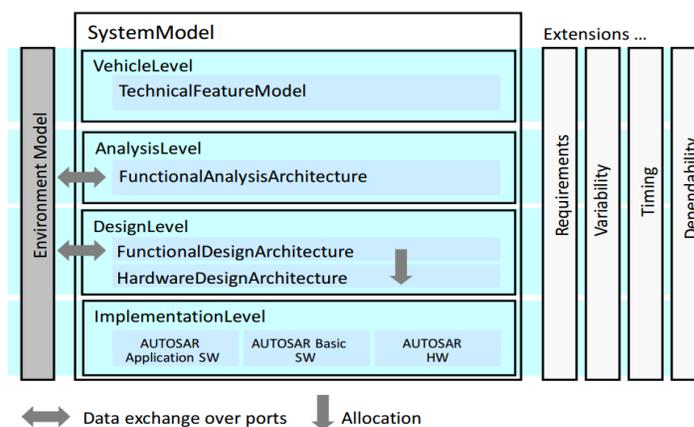


Figure 7. The *EAST-ADL* abstraction layers [7]

*EAST-ADL* covers all phases of our taxonomy, as we make use of its level system. Nevertheless, it is discussable, whether the implementation level is also covered, because it embeds *AUTOSAR* instead of a distinct solution. In the view of the authors, this can be ignored, because many methods imply predefined languages (which omits extensive redevelopment). *EAST-ADL* includes lines of action for the implementation level. This can be seen as part of a major surrounding process. Other phases are performed by use of own languages or language aspects, which can be acknowledged as *DSL*. A tool is not comprised, though some implementations are available.

*EAST-ADL* contains many information directly related to the automotive domain. The integration and use is therefore easy, whereas the lack of a proper implementation or tool for many years prevented the distribution in the domain. In line with this, none of the respondents of our survey uses *EAST-ADL* and it is scarcely known [9].

#### 5.4. MATLAB/Simulink/TargetLink

*MathWorks MATLAB/Simulink* and *dSPACE TargetLink* [17],[28] compose a software modeling framework used to create a software model and its derived target code. *Simulink* is a graphical data flow modeling language embedded in the *MATLAB* computing environment. Models created in *Simulink* consist of so-called *blocks* (functional entities), which can be linked to each other and are taken out of a predefined block library. The models are closely related to the hardware structure, which also becomes apparent in the type of blocks available in the library, e.g. bus-, mux-/demux or gain-blocks. *TargetLink* provides target source code generation out of the created models. Testing, verification and validation methods are also available.

The method is started at the *Detailed Design* phase and continues until the corresponding *Integration Test*. *MATLAB* is the basic tool framework. It is mandatory for the use of *Simulink*, a graphical *DSL* used to create the required models. *TargetLink* is used to create target source code out of the models. No lines of action are included.

The *MATLAB/Simulink/TargetLink*-tool chain is one of the major software engineering frameworks currently used in the automotive domain. This is also illustrated in our survey, where at least two-third of the respondents already use the tool chain and more than 86% are familiar with it. However, it mainly lacks possibilities to design the system architecture or to include requirements at an abstract level. Consequently, the system engineering in this case is rather bottom-up and implementation-related instead of being top-down and iterative as required by the *V-Model*.

#### 5.5. SCADE

The *Esterel Safety Critical Application Design Environment (SCADE)* [20] is a software development framework initially grounded in the avionics industry. It consists of four different tools, whereof *SCADE Suite* is focused on model-based software development. As basis, the formal, synchronous and data flow-oriented *DSL Lustre* [15] is used, which utilizes graphical models to describe the system-underdevelopment. The *SCADE Suite* includes methods for validation/verification and code generation.

The *SCADE*-tool chain covers the complete *V-Model*, so all phases of the taxonomy are included. With *Lustre*, a *DSL* is used. *SCADE* contains detailed process information and lines of action.

*SCADE* is well-known in its initial application area, the avionic industry, but has recently been introduced to first automotive projects. Still, an adoption for this new context requires a certain amount of modifications, e.g. the introduction of automotive-related concepts and definitions. In our survey, none of the respondents practically uses *SCADE*, whereas one-fifth are at least aware of this method.

## 5.6. ADTF

*Automotive Data and Time-Triggered Framework (ADTF)* [19] is a software modeling framework aiming at the development of driver-assistance features. *ADTF* allows real-time data playback and provides visualization features that are used to simulate the created models and evaluate it according to defined timing constraints. This guarantees, that both the simulation on the development system and on the target system act and react similar. The *ADTF*-models consist of graphical representations of functions, so-called *filters*, with their inputs and outputs (e.g. signals). As data source, different standardized sources like CAN or camera data can be used simultaneously and synchronized.

*ADTF* is a tool with focus on the development of car functions. It ranges from the *Analysis Level* until the *Implementation Level* with the integration of production code. Testing is limited to simple manual tests. Lines of action are not included, whereas the models are created with help of a graphical *DSL*. The functional range lacks detailed architecture and testing features.

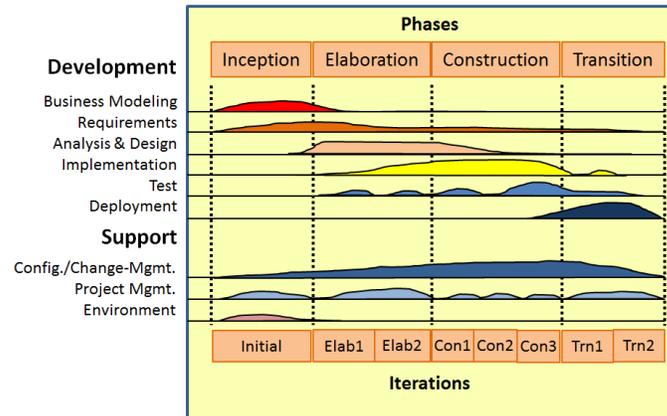
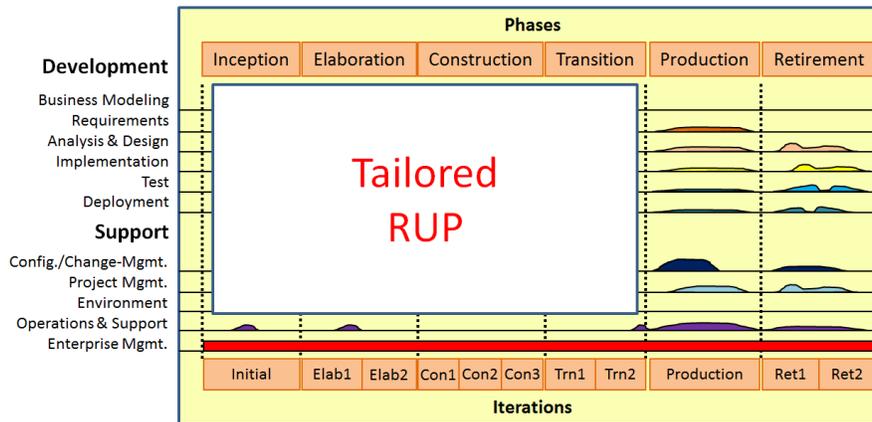
*ADTF* was initially developed for the automotive domain in Germany in 2011. This, in conjunction with our survey being carried out in the environment of German car manufacturers and their suppliers explain the high familiarity of the respondents with *ADTF* and the utilization rate of 50% [9]. In foreign markets, this rate would be much lower. Hence, the use of *ADTF* is limited so far to German car manufacturers.

## 5.7. RUP/EUP

The *IBM Rational Unified Process (RUP)* [1],[27] is an iterative software development process. It is split into four phases that handle the project definition, system architecture, implementation and delivery. Each phase contains a set of engineering disciplines, which may occur iteratively. Beside the general process model, *RUP* contains best practices, templates and checklists to support the developer. The complete process setup and the importance of each discipline for each particular phase is shown in figure 8, at which the ordinate indicates the required time and effort at a specific time.

An enhancement to *RUP* is proposed as *Enterprise Unified Process (EUP)* [1]. It adds two new phases, that handle maintenance and retirement. Additionally, new disciplines are added (cf. figure 9). The intention is to cover the more generic and development-independent topics like personnel administration.

The development ranges from the specification to the retirement of the finished product, so all levels of the taxonomy are covered. According to [1], both methods are processes with no integrated languages or tools. To make use of them, a separate implementation is required which is not part of the original definition. Anyway, work flows and process steps can be adopted for given project scenarios.

Figure 8. The *RUP* phases and disciplines [1],[27]Figure 9. The *EUP* phases and disciplines [1],[27]

Both *RUP* and *EUP* are primarily general processes without an implementation or any automotive focus, so the practical use in the automotive context is rather limited. Our survey states, that none of the respondents actually uses *RUP/EUP* and only a minor part is familiar with them [9].

## 5.8. SimTAny

*Simulation and Test of Anything (SimTAny)* [16] (formerly known as *VeriTAS*) represents a framework that provides the *test-driven agile simulation (TAS)* process and a respective tool chain. The process specifies, that the system and the usage model are derived separately from the requirements and specified by individual *UML* models (see figure 10). A respective simulation model is automatically generated from the system model and test cases are automatically derived from the usage model. Subsequently, the simulation model is run together with test cases in a simulation. An implementation of the system or the hardware is not required. Thus, it is possible to identify modeling errors or inconsistencies in the system model and/or the usage model and validate them early in the development process.

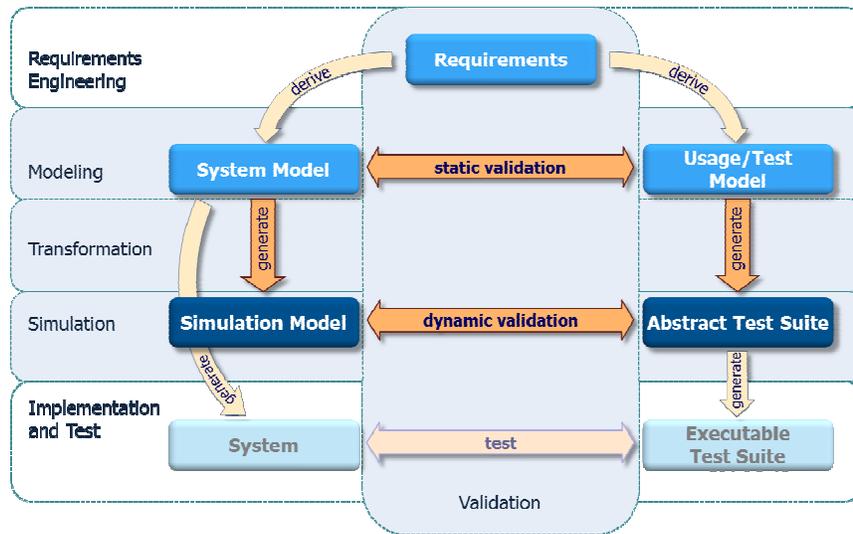


Figure 10. Test-driven agile simulation process provided by *SimTAny* [34]

Although the integration of requirements is possible, it is not the focus of *SimTAny*. Therefore the Analysis Level is the actual starting point. *SimTAny* places emphasis on the simulation of the system to be developed and does not include production code. So the *Implementation Level* and all subsequent levels are not covered. A surrounding process and a method implementation are part of *SimTAny*. As specification language, *UML* as single *GPL* is used.

*SimTAny* is mainly applied by academics or in research and therefore not used in the automotive domain so far. First projects to introduce it to the domain are currently running. Unsurprisingly, in our survey, only respondents located in research already work with *SimTAny* [9].

## 6. CONCLUSION AND FUTURE WORK

Selecting a software engineering method, that satisfies the requirements of an automotive project, is a difficult task. In order to aid the decision making, a well-structured overview, as well as a possibility to compare the features of the available approaches are required. There exist several taxonomies that provide such an overview, however, they mainly lack the automotive focus or are restricted to a specific software engineering method type. As outlined in the introduction, such an overview can be necessary for a development decision in a given project scenario, even if the investigated methods differ extensively.

That is why this paper outlines a new taxonomy for software engineering approaches focused on the automotive domain. It consists of a combination of the general *V-Model*, the level model taken out of *EAST-ADL* and the enrichment with the indication, whether *GPLs* or *DLSs* are included. Due to clarity and simplicity reasons, the results are depicted in a diagram (cf. figure 4). This allows the reader to easily compare several possibly quite different engineering approaches.

The introduced taxonomy has been applied to currently established key-methods in the domain. The result is a well-structured overview that serves as a compendium and exemplifies the approach. This approach has been reviewed in our survey by the respondents to get an indicator, how helpful, self-explanatory and useful this taxonomy appears to the target user group. The

resulting evaluation values  $e = 6.36$  with  $e \in [1, 9]$ , 1 as representative for *not helpful* and 9 for *helpful*. This is sufficient to state the taxonomy as helpful, though this value can be increased by adding more information to the taxonomy or applying it to more different methods to provide a diversified information base for project decisions.

The taxonomy approach described in this paper is the first step in the development of a detailed classification pattern for software engineering methods in the automotive domain. The proposed format and diagram can be prospectively enriched with more classification information or can be extended with new phases/levels. As depicted in our survey, there are several additional characteristics of engineering methods that are more or less important for engineers [9]:

- important: support, extensibility, documentation, training courses
- neutral: amount of features, market share, price
- unimportant: familiarity of the manufacturer

These values cannot be linked with all types of engineering methods, e.g. processes partly have no manufacturers. Instead they are defined by standardization organizations. As a result, this list of characteristics is not yet included in our taxonomy. There are two ways of incorporating these values into the decision process. First, the values can be included by taking a subset of characteristics, that is matchable to the investigated methods and enriching the taxonomy with this subset. Second, our taxonomy can be used to constrain the list of investigated methods and afterwards, other taxonomies (e.g. developed by Broy [14]) can be used in combination with the whole set of characteristics to determine a final solution for the given project scenario. In both cases, our taxonomy serves as first easy-to-use decision guidance.

## REFERENCES

- [1] Ambler, S. and Nalbone, J. and Vizdos, M.: The Enterprise Unified Process. Prentice Hall Press, Upper Saddle River, NJ, USA (2005)
- [2] AUTOSAR development partnership: AUTOSAR, <http://www.autosar.org>, accessed 02-November-2015
- [3] AUTOSAR GbR: AUTOSAR Layered Software Architecture (March 2006)
- [4] AUTOSAR Tool Platform User Group: ARTText – An AUTOSAR Textual Language Framework, <http://www.artop.org/artext/>, accessed 02-November-2015
- [5] Babar, M.A. and Gorton, I.: Comparison of scenario-based software architecture evaluation methods. In: 11th Asia-Pacific Software Engineering Conference, 2004. pp. 600-607. APSEC '04 (2004)
- [6] Bangia, R.: Dictionary of Information Technology. Laxmi Publications Ltd. (2010)
- [7] Blom, H. and Hagl, F. and Papadopoulos, Y. and Reiser, M.-O. and Sjöstedt, C.-J. and Chen, D.-J. and Kolagari, R.T.: EAST-ADL - An Architecture Description Language for Automotive Software-Intensive Systems. International Standard (2012)
- [8] Blum, B.I.: A Taxonomy of Software Development Methods. Communications of the ACM 37(11), 82-94 (1994)

- [9] Bock, F.: Survey: Software Engineering Methods in the Automotive Domain (2015), Raw data available at [http://www7content.cs.fau.de/%7Ebock/2015\\_10\\_bock\\_\\_raw\\_data.zip](http://www7content.cs.fau.de/%7Ebock/2015_10_bock__raw_data.zip), accessed 02-November-2015
- [10] Braun P. and Broy, M. and Houdek, F. and Kirchmayr, M. and Müller, M. and Penzenstadler, B. and Pohl, K. and Weyer, T.: Guiding requirements engineering for software-intensive embedded systems in the automotive industry. *Computer Science - R&D* 29(1), 21-43 (2014)
- [11] Bringmann, E. and Krämer, A.: Model-Based Testing of Automotive Systems. In: *Proceedings of the 2008 International Conference on Software Testing, Verification, and Validation*. pp. 485-493. ICST '08, Washington, DC, USA (2008)
- [12] Bröhl, A.P.: *The V-Model. Software – Application Development - Information Systems* (in German), Oldenbourg, Munich (1993)
- [13] Broy, M.: Challenges in Automotive Software Engineering. In: *Proceedings of the 28th International Conference on Software Engineering*. pp. 33-42. ICSE '06, New York, NY, USA (2006)
- [14] Broy, M. and Feilkas, M. and Herrmannsdörfer, M. and Merenda, S. and Ratiu, D.: Seamless Model-Based Development: From Isolated Tools to Integrated Model Engineering Environments. *Proceedings of the IEEE* 98(4), 526-545 (2010)
- [15] Caspi, P. and Pilaud, D. and Halbwachs, N. and Plaice, J. A.: LUSTRE: A Declarative Language for Real-time Programming. In: *Proceedings of the 14th ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages*. pp. 178-188. POPL '87, New York, NY, USA (1987)
- [16] Djanatljev, A. and Dulz, W. and German, R. and Schneider, V.: Veritas - A Versatile Modeling Environment for Test-Driven Agile Simulation. In: *Proceedings of the 2011 Winter Simulation Conference*. WSC 2011, Phoenix, AZ, USA (2011)
- [17] dSpace: TargetLink, <http://www.dspace.com/en/pub/home/products/sw/pcgs/targetli.cfm>, accessed 02-November-2015
- [18] EAST-ADL Association: EAST-ADL Domain Model Specification (2013)
- [19] Elektrobit: EB Assist ADTF - Driver assistance systems start with EB Assist ADTF, <https://automotive.elektrobit.com/products/eb-assist/adtf/>, accessed 02-November-2015
- [20] Esterel: SCADE Suite Control Software Design, <http://www.esterel-technologies.com/products/scade-suite/>, accessed 02-November-2015
- [21] Fürst, S. and Mössinger, J. and Bunzel, S. and Weber, T. and Kirschke-Biller, F. and Heitkämper, P. and Kinkelin, G. and Nishikawa, K. and Lange, K.: AUTOSAR - A Worldwide Standard is on the Road, <http://www.win.tue.nl/~mvdbrand/courses/sse/0809/papers/AUTOSAR.pdf>, unpublished report
- [22] Hoffmann, H.P.: *Systems Engineering Best Practices with the Rational Solution for Systems and Software Engineering Deskbook Release 4.1. Manual* (2014)
- [23] Hofmeister, C. and Kruchten, P. and Nord, R.L. and Obbink, H. and Ran, A. and America, P.: Generalizing a Model of Software Architecture Design from Five Industrial Approaches. In: *5th Working IEEE/IFIP Conference on Software Architecture*, 2005. pp. 77-88. WICSA 2005 (2005)
- [24] IEEE: IEEE Standards Definition Database, <http://dictionary.ieee.org>, accessed 02-November-2015

- [25] Kitchenham, B.A.: Evaluating Software Engineering Methods and Tool Part 1: The Evaluation Context and Evaluation Methods. SIGSOFT Software Engineering Notes 21(1), 11-14 (1996)
- [26] Kosar, T. and Oliveira, N. and Mernik, M. and Pereira, M.J.V. and Črepinšek, M. and da Cruz, D. and Henriques, P.R.: Comparing General-Purpose and Domain-Specific Languages: An Empirical Study. *Computer Science and Information Systems* 7(2), 247-264 (2010)
- [27] Kruchten, P.: *The Rational Unified Process: An Introduction*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (2003)
- [28] MathWorks: Simulink - Simulation and Model-Based Design, <http://www.mathworks.com/products/simulink/>, accessed 02-November-2015
- [29] Naur, P. and Randell, B. (ed.): *Software Engineering: Report of a Conference Sponsored by the NATO Science Committee, Garmisch, Germany, 7-11 Oct. 1968*, Brussels, Scientific Affairs Division, NATO. NATO Science Committee (1969)
- [30] Object Management Group: SysML Open Source Specification Project, <http://www.sysml.org>, Standard, accessed 02-November-2015
- [31] Object Management Group: Unified Modeling Language (UML) Resource Page, <http://www.uml.org>, Standard, accessed 02-November-2015
- [32] Oliveira, N. and Pereira, M.J.V. and Henriques, P.R. and da Cruz, D.: Domain-Specific Languages – A Theoretical Survey. In: *Proceedings of the 3rd Compilers, Programming Languages, Related Technologies and Applications*. pp. 35-46. CORTA '2009 (2009)
- [33] Sbadka, D.: Impacts of shortening Product Life Cycle in the Automotive Industry. *Transfer inovácií* 29/2013 (2013)
- [34] Schneider, V. and German, R.: Integration of Test-driven Agile Simulation Approach in Service-oriented Tool Environment. In: *Proceedings of the 46th Annual Simulation Symposium*. pp. 11:1-11:7. ANSS 2013, San Diego, CA, USA (2013)
- [35] Sun, Y. and Demirezen, Z. and Mernik, M. and Gray, J. and Bryant, B.: Is My DSL a Modeling or Programming Language? In: *Domain-Specific Program Development*. p. 4. Nashville, TN, USA (2008)

# APPLICATION OF BICLUSTERING TECHNIQUE IN MACHINE MONITORING

Marcin Michalak

Institute of Informatics, Silesian University of Technology, Gliwice, Poland

Marcin.Michalak@polsl.pl

## **ABSTRACT**

*Nowadays we can observe the change of the structure of energy resources, which leads to the increasing fraction of a renewable energy sources. Traditional underground coal mining loses its significance in a total but there are countries, including Poland, which economy is still coal based. A decreasing coal resources imply an exploitation a becoming harder accessible coal beds what is connected with the increase of the safety of the operation. One of the most important technical factor of the safety of underground coal mining is the diagnostic state of a longwall powered roof support. It consists of dozen (or hundreds) of units working in a row. The diagnostic state of a powered roof supports depends on the diagnostic state of all units. This paper describes the possibility of unit diagnostic state analysis based on the biclustering methods.*

## **KEYWORDS**

*Biclustering, Machine Monitoring, Machine Diagnosis*

## **1. INTRODUCTION**

In a coal mining industry – similarly as in the case of other industry branches – the growth of monitoring systems application. Initially, monitoring systems were designed just for the purpose of data acquisition and presentation is being observed. Over time, their abilities were extended in the direction of simple dangerous situations recognition and finally – to the advanced machine diagnostic state analysis and its prediction for the nearest future.

Longwall systems are the basis of the coal mining, because the longwall is the place in the process of mining from which we can say about the output. Mechanised longwall systems consist of longwall shearer (which tears off the output from the rock), longwall conveyor (transports the output from the longwall to the heading) and units of powered roof support (prop the roof after mining the output).

Longwall systems are very interesting objects from the collected data point of view. Its most important part is a power roof support. Its primary task is to protect the other elements of the longwall system, especially the coal shearer which is an essential part as of a coal mining process, not to mention the protection the workers from the falling rocks. Power roof support consists of units. In a particular moments of time – after shearer takes the another part of the

longwall output – each unit has to move in the direction of the whole longwall face advance (treading), protecting the rock material, exposed by the shearer, from collapsing.

Unequal propping can be caused by leaks in the hydraulic system (pipes, valves) or leaks in legs of the unit. To long times of treading can point the wrong diagnostic state of the unit or be caused by the wrong usage (so called: moving with the contact of roof-bar with the roof). It is also dangerous to perform the treading too long as the roof is not propped. So it can be stated that the safety of coal mining is determined by the diagnostic state of all parts of the longwall mining system, including the diagnostic state of all units of power roof support.

In this article the ability of adaptation of biclustering method for the purpose of the analysis the data from longwall monitoring systems is presented. The paper is organized as follows: it starts from the brief description of monitoring systems with a special consideration of underground coal mining monitoring system is presented. Then the description of a construction, the role and a working cycle of a powered roof support is presented together with the proposed monitored (and extracted) parameters. Next part presents a wide group om continuous and binary data biclustering methods. The paper ends with application of an OPSM biclustering algorithm for the real data.

## 2. MONITORING SYSTEMS

Nowadays, software producers and monitoring systems users point the need of analysis of the data, collected in repositories of these systems. In particular, the definition of diagnostic models of monitored devices can be a goal of this analysis [8]. The process of a diagnostic model identification can be carried out by planned experiments or on the basis of a data from the past device operation. In a specific situation, when there are no data describing an improper state of the machine it is possible to develop a model describing only the proper state of the machine and treat the deviation from the model as the possibility of improper machine state [11].

The problem of a monitoring and diagnosing of a coal mine industry devices was raised recently in [1][4][7][11][12][17]. These topics are presented widely and review in [23]. In these works also new methods of extraction and processing of new diagnostic features in new diagnostic relations discovering are presented. Especially in [1] the diagnostic of conveyor belts is described. In [18] a current consumption and a temperature of roadheader cutting head. On the basis of these parameters three roadheader working states were defined. Two of them described different but correct underground mining conditions. In this paper also the parameter reflecting the roadheader cooling system efficiency was defined.

Longwall conveyors diagnostic was an aim of the following works [4][7][12]. In [4]the way of conveyor chute failures detection on the bottom side of a conveyor was presented. On the basis of the conveyor engines power consumption analysis the failure was detected with an accuracy to the one unit. In [7] the complex subassemblies management system was proposed, which allows to generate operational and analytical reports as well as summary statements. In [12] a harmonic analysis was used for the prediction of a diagnostic state of longwall conveyor chain.

### 3. BICLUSTERING

Biclustering is the problem of unsupervised data analysis, where we are grouping scalars from the two-dimensional matrix. It called also as co-clustering, two-dimensional clustering or twomode clustering. This approach has been started in 70's in the last century [6] and is successfully applied in bioinformatics [3][15][21]. The idea of biclustering is to find in a matrix a subset of columns and subset of rows, which intersection gives a submatrix of cells with similar (or the same) values.

In the literature there are a lot of algorithms of biclusters induction. In [3] authors define bicluster as a subset of rows under subset of columns, for which calculated parameter (mean squared residue score) is below threshold defined by the user. The minimum value of the considered parameter is 0. The algorithm consists of two steps. Initially the rows and columns are removed from input dataset, until the value of mean squared residue score is below assumed level. Then rows and columns, which were removed during the first step, are added to obtained in the previous step submatrix until its score fulfils the criterion of being bicluster. After each iteration, the founded bicluster has been hidden with random values. The extension of this algorithm proposed in [22] allows to avoid noise among input dataset, which was a consequence of masking discovered biclusters.

The Order Preserving Submatrix Algorithm was presented in [2]. The bicluster was defined as subset of rows, which preserves linear ordering across subset of conditions. The set of valid biclusters is identifying by algorithm based on stochastic model. This idea was also evolved in [9].

The algorithm X-Motif is dedicated to the extraction of conserved gene expression motifs from gene expression data and has been proposed in [14]. Bicluster is defined as subset of genes, which expression level belong to the same state across subset of conditions. The states are assigned to genes during preprocessing step. In order to find multiple biclusters an algorithm is running in an iterative way. Each iteration starts from different initial sets.

There exists also methods of biclustering dedicated for matrices with the binary values. Bimax [16] uses a simple divide-and-conquer approach for finding all inclusive maximal biclusters for a given minimal number of rows and columns. Bicluster, which is maximal in the sense of inclusion is defined as not entirely contained in any other bicluster. Such assumption allows to exclude from analysis individual cells equal to one, which can be considered as a single biclusters, however they provide no important information.

BicBin [20] is an algorithm dedicated for binary sparse matrices. It consists of three main components: the score function to evaluate a submatrix, the search algorithm to restrict the search space of all possible submatrices and an algorithm used to extract all biclusters in an ordered way from a dataset. BicBin is dedicated for finding inexact biclusters. Each run of BicBin may give different results, because algorithm finds set of random biclusters, which fulfil its restrictions and cover all ones in dataset.

A novel approach of the binary matrix biclustering is based on the rough sets theory [13] where non-exact biclusters are defined as the ordered pair of biclusters called a lower and an upper approximation. The lower approximation is the exact submatrix of the given one and the upper

approximation is non-exact matrix that is the superset of a given one. The algorithm is hierarchic similarly as the Ward clustering algorithm and it starts from the previously generated set of exact biclusters [19]. In every step two rough biclusters can be joined if the intersection of their lower approximations is nonempty. The generalisation of the data description allows to limit the number of final biclusters assuring the assumed level of the description accuracy.

The analogical hierarchical strategy can be also applied for classical biclusters (not considered as the rough bicluster) and was presented in [10].

#### 4. POWERED ROOF SUPPORT

The final safety of operations in a coal mine is dependent on several components: a human factor, natural influence, a technical reasons and – of course – from the unexpected circumstances. A technical operating conditions are an object of interest of multiple monitoring systems [7]. The most common monitored signal is the pressure in the leg (legs) of the section, which reflects the real strength of propping the roof.

A single section of a powered roof support consists of one or more hydraulic prop (legs), which holds up an upper part of the section (roof-bar). A section has also and hydraulic shifting system, which is responsible for shifting the unit with the longwall advance simultaneously. Most of the time the unit props the roof, assuring the safety of mining, but after each shearer run it moves to prop the newly bared roof. A single unit is presented on the Fig. 1.



Fig. 1 Single unit of powered roof support ([www.joy.com](http://www.joy.com)).

A typical unit working cycle will be presented on the real – 6000 seconds long – observation of the pressure in one of the two-leg unit, shown on the Fig. 2. As it can be observed, it is common situation when two legs are propping the roof-bar with different force (they have different pressure levels in legs).

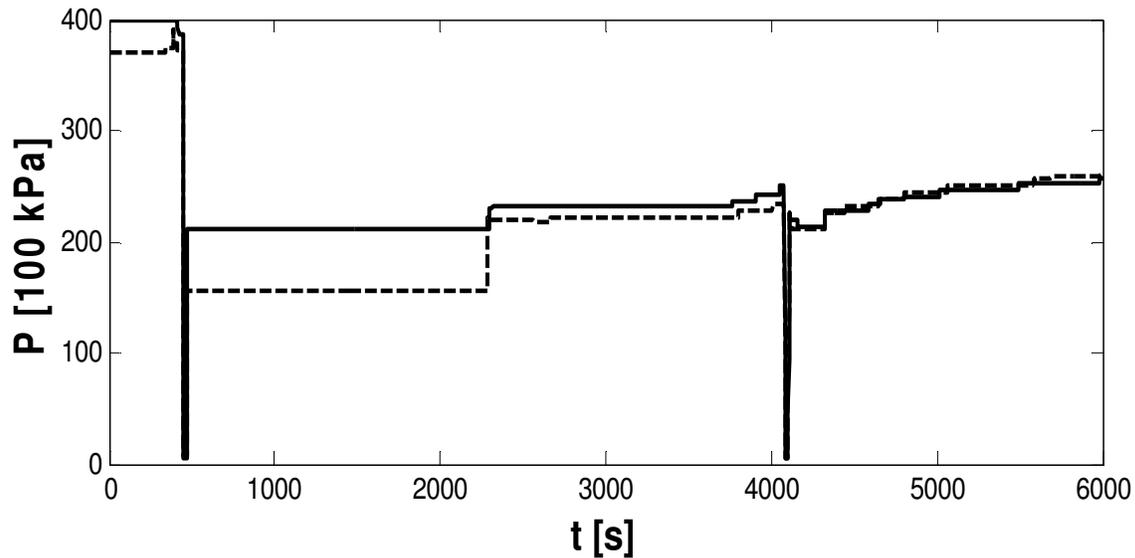


Fig. 2 A real time series of pressure in two legs of the unit.

A typical powered roof support unit working cycle can be divided into several phases. Starting from the moment of the beginning of unit shifting the following phases are named and described as follows:

- treading,
- spragging,
- overbuilding,
- pre-treading,
- pressure lowering.

For each of these phases some similar and some specific variables can be defined, for example: phase duration (for all phases), pressure increase speed (spragging, overbuilding), pre-treading dynamic type, pressure level equability.

The analysis of the mentioned phase statistics requires a diagnostic information, or a diagnostic algorithm, that classified the moment of time (a value of a pressure) into a proper working cycle phase. Otherwise, it is not possible to analyse the whole data – each column representing a single second of powered roof support section work. Then some aggregation of the data must be performed.

The time series, describing a single powered roof support unit that contains several legs, can be an average pressure level in legs or a difference between leg pressure levels. As an aggregate a multiple statistics can be taken into considerations, starting from a mean, maximal and minimal value. The way of data aggregation and a time of aggregation should be dependent on the goal of analysis. For example: using a minimum value in the aggregation time will help to detect a time slices when the section was in a treading phase.

## 5. EXPERIMENTS

In this paper an application of biclustering methods for pressure level equability in legs of one powered roof support unit will be presented. Application of these methods will help to find a subset of sections which behave similarly in a selected period (or periods – the time interval does not have to be connected). On the basis of a real data, a procedure of diagnostic procedure will be presented. The real data describes the over 8 hours of over 100 sections of powered roof support work. Each section was a double-leg and a pressure level measurement was taken once in a second. On the Fig. 3 an average pressure of two legs is presented. Small values of the absolute difference are marked as white points. As the absolute difference increases the point color becomes more black.

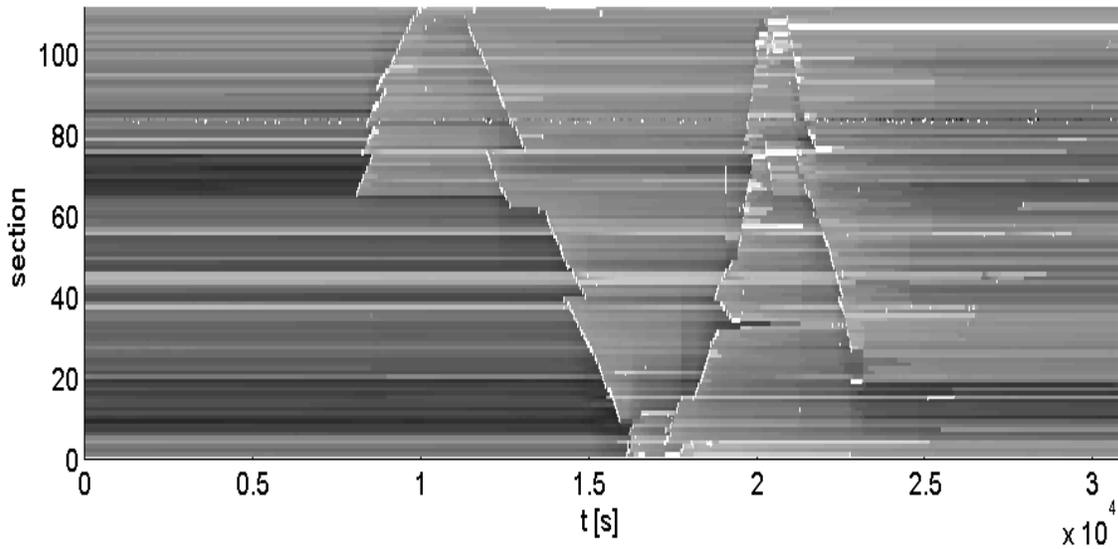


Fig. 3 Average pressure in two legs.

The horizontal zigzag, starting at the 150th minute and ending at the 450th minute refers to the shearer passage and the sections treading. As it can be seen, each section was moved several times. Such small number of treadings does not allow to analyse phases durations. Instead, the pressure level inequality will become the goal of the analysis.

The relative difference of a pressure levels was calculated with the following formula:

$$avg = \frac{|leg_1 - leg_2|}{\min\{leg_1, leg_2\}}$$

where  $leg_1$  and  $leg_2$  are pressure levels in according legs.

Due to the large amount of the data and a goal of data generalisation the aggregation time is set to 60 seconds. As the point of the interest is a difference of a pressure levels the aggregation measure is maximal value of differences in one minute.

The definition of a difference for the pressure level allows to have infinities, as the division by zero is not forbidden. This situation occurs when one of the legs is not propping or there was an error in a measurement device or in a data transmission. The infinity can be easily replaced by 750, which is higher than the highest absolute difference between legs pressure (equal to 70.4 [MPa]). The interpretation of this replacement can be as follows: as only one leg props there is a maximal inequality.

For the further analysis only a subset of sections and a period of time is selected: it will contain 51 middle sections monitored during the middle 200 minutes (Fig. 4).

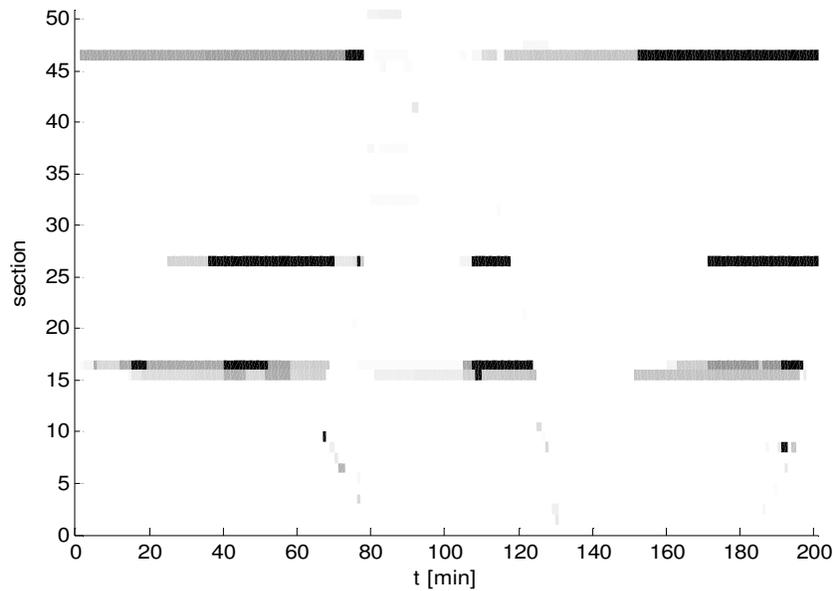


Fig. 4 An aggregated maximal differences of pressure levels in units.

A statistical description of the subset values is presented in the Table 1. The bicluster 0 means the whole data while the following ones are the selected biclusters generated with OPSM algorithm [5]. With the standard algorithm settings it generated 42 biclusters. Four of them are briefly described in the Table 1 and also presented on the Fig. 5. We can observe that the algorithm detected biclusters with the very high pressure level inequality.

Table 1. Statistical description of real data and selected of generated biclusters.

#	min	Q1	Q2	Q3	mean	max	std	duration [min]	sections
0	0.000	0.036	0.085	0.200	20.278	750	103.823	201	51
1	0.000	0.045	0.683	122.286	159.000	750	228.400	201	2
3	0.000	0.053	0.372	85.428	9.667	750	197.493	182	4
5	0.004	0.091	0.152	74.451	2.225	750	194.031	127	6
11	0.020	0.100	0.157	62.837	0.386	750	207.416	38	12

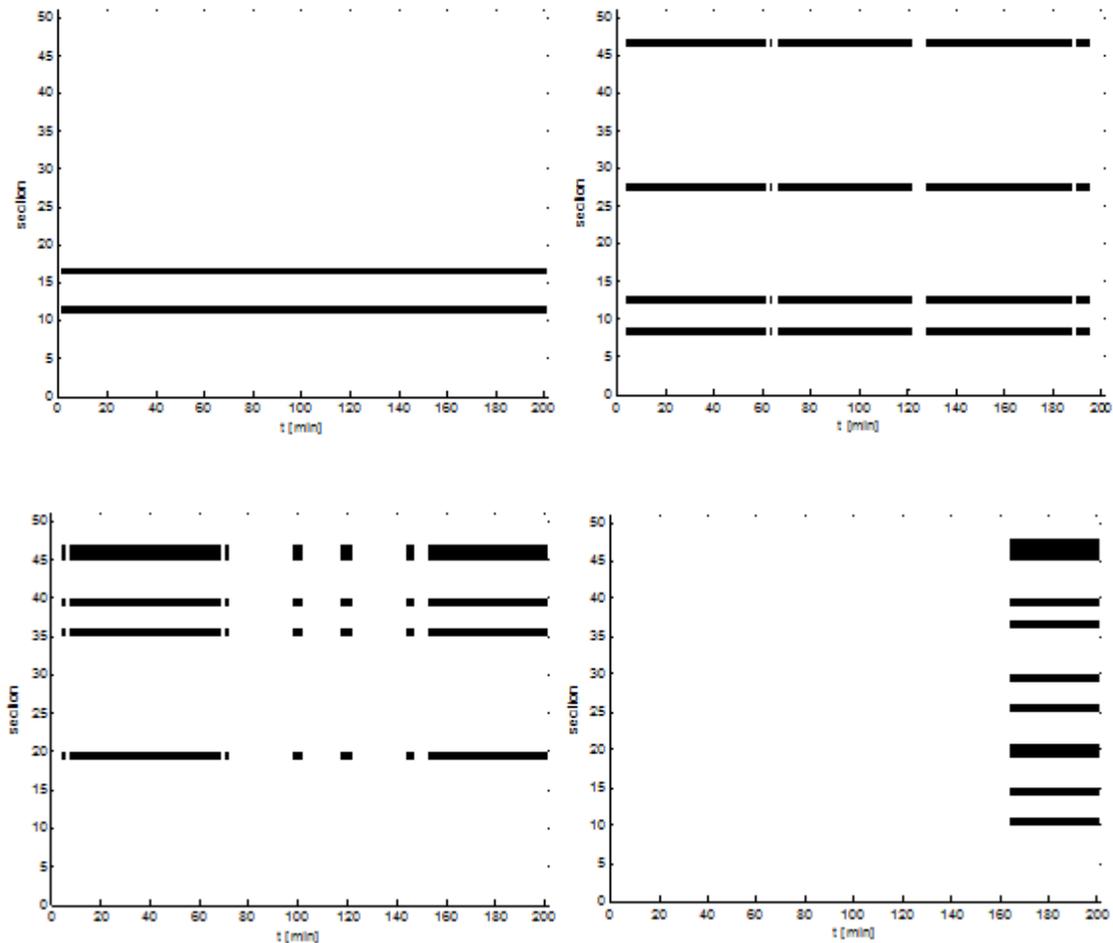


Fig. 5 Four biclusters: 1st (upper left), 3rd (upper right), 5th (lower left) and 11th (lower right).

## 5. CONCLUSIONS

In the paper the application of biclustering technique for the off-line machine monitoring was presented. This simple example shows a possibility of building procedures for the purpose of a variety aspects of powered roof support monitoring and diagnosing. The presented approach does not exhaust the subject of finding biclusters in a longwall complex data. Future works in this area will focus on joining biclustering results with real improprieties in the real data.

## ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science and Higher Education – internal grant signature: BKM 515/RAu2/2015.

**REFERENCES**

- [1] Bartelmus W.: Condition Monitoring of Open Cast Mining Machinery. Wroclaw University of Technology Press, Wroclaw 2006.
- [2] Ben-Dor A., Chor B., Karp R., Yakhini Z.: Discovering Local Structure in Gene Expression Data: The Order-Preserving Sub-Matrix Problem, Proceedings of the 6th Annual International Conference on Computational Biology, 2002
- [3] Cheng Y., Church G.: Biclustering of Expression Data, Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology, 93-103, 2000.
- [4] Gąsior S.: Diagnosis of longwall chain conveyor, *Przegląd Górniczy*, 57(7-8):33-36, 2001.
- [5] Gupta J., Singh S., Verma N.: MTBA: MATLAB Toolbox for Biclustering Analysis, IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions, IIT Kanpur, India, pp. 94-97, July 2013.
- [6] Hartigan J.A.: Direct Clustering of a Data Matrix. *Journal of American Statistical Association*, 67(337):123-129, 1972
- [7] Kacprzak M., Kulinowski P., Wędrychowicz D.: Computerized information system used for management of mining belt conveyors operation. *Eksploracja i Niezawodność - Maintenance and Reliability*, 13(2):81-93, 2011.
- [8] Korbicz J., Kościelny J.M., Kowalczyk Z., Cholewa W.: *Fault Diagnosis: Models, Artificial Intelligence, Applications*, Springer, 2004.
- [9] Liu J., Wang W.: OP-Clusters: Clustering by Tendency in High Dimensional Space, Proceedings of the 3rd IEEE Int. Conf. on Data Mining, 187-194, 2003.
- [10] Michalak M., Lachor M., Polański A.: HiBi - The Algorithm of Biclustering the Discrete Data. *Lecture Notes in Computer Science* 8468:760-771, 2014.
- [11] Michalak M., Sikora B., Sobczyk J.: Diagnostic Model for Longwall Conveyor Engines, *Advances in Intelligence and Soft Computing*, 391:437-447.
- [12] Michalak M., Sikora M., Sobczyk J.: Analysis of the longwall conveyor chain based on a harmonic analysis. *Eksploracja i Niezawodność - Maintenance and Reliability* 2013; 15 (4): 332-336.
- [13] Michalak M., Stawarz M.: HRoBi - The Algorithm for Hierarchical Rough Biclustering, *Lecture Notes in Computer Science*, 7895:194-205, 2013.
- [14] Murali T.M., Kasif S.: Extracting Conserved Gene Expression Motifs from Gene Expression Data, *Pacific Symposium on Biocomputing*, 77-88, 2003.
- [15] Pensa R., Boulicaut J.F.: Constrained Co-clustering of Gene Expression Data, *Proc. SIAM International Conference on Data Mining, SDM 2008*, 25-36, 2008
- [16] Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Grissens W., Hennig L., Thiele L., Zitzler E.: A Systematic Comparison and Evaluation of Biclustering Methods for Gene Expression Data, *Bioinformatics* 22(9):1122-1129, 2006.

- [17] Sikora M.: Induction and pruning of classification rules for prediction of microseismic hazards in coal mines. *Expert Systems with Applications* 38(6):6748-6758, 2011.
- [18] Sikora M, Michalak M.: Eksploracja baz danych systemów monitorowania na przykładzie obserwacji pracy kombajnu chodnikowego, *Bazy Danych: Rozwój metod i technologii. Tom I*, (in Polish), WK L, Warsaw 2008:429-437.
- [19] Stawarz M., Michalak M.: eBi — The Algorithm for Exact Biclustering. *Lecture Notes in Computer Science* 7268:327-334, 2012.
- [20] van Uitert M., Meuleman W., Wessels L.: Biclustering Sparse Binary Genomic Data, *Journal of Computational Biology*, 15(10):1329-1345, 2008.
- [21] Yang E. Foteinou P.T., King K.R., Yarmush M.L., Androulakis I.P.: A Novel Nonoverlapping bislustering Algorithm for Network Generation Using Living Cell Array Data, *Bioinformatics* 17(23) 2306-2313, 2007.
- [22] Yang J., Wang H., Wang W., Yu P.: Enhanced biclustering on expression data, *Third IEEE Symposium on Bioinformatics and Bioengineering*, 321-327, 2003.
- [23] Zimroz R.: *Metody adaptacyjne w diagnostyce układów napędowych maszyn górniczych* (in Polish). Wrocław University of Technology Press, Wrocław 2010.

## AUTHOR

Marcin Michalak Marcin Michalak was born in Poland in 1981. He received his M.Sc. Eng. In computer science from the Silesian University of Technology in 2005 and Ph.D. degree in 2009 from the same university. His scientific interests are in machine learning, data mining, rough sets and biclustering. He is an author and coauthor of over 60 scientific papers.



# VIRTUAL SCENE CONSTRUCTION OF LARGE-SCALE CULTURAL HERITAGE : A FRAMEWORK INITIATED FROM THE CASE STUDY OF THE GRAND CANAL OF CHINA

Jian Tan<sup>1</sup> and Shenghua Wang<sup>2</sup>

<sup>1</sup>Key Laboratory of Digital Earth Science, Institute of Remote Sensing and  
Digital Earth, Chinese Academy of Sciences, Beijing, China

Tanjian1998@gmail.com

<sup>2</sup>School of Public Administration and Communication, Beijing Information  
Science Technology University, Beijing, China

386974698@qq.com

## **ABSTRACT**

*Virtual reality technology has been applied to the protection of cultural heritage for about 20 years. However, methods or systems of cultural heritage reported in previous studies are still unable to represent large-scale cultural heritage sites such as the Beijing-Hangzhou Grand Canal, the Struve Geodetic Arc and the boundaries of the Roman Empire. We aimed at constructing a large-scale cultural-heritage 3-D model with the focus on better management and organization of the scene. Starting from the case study of the Beijing-Hangzhou Grand Canal, we first explore various remote sensing data suitable for large-scale cultural heritage modeling, and then adopt a 3-D geographic global information system for large-scale 3-D scene organization and management.*

*The entire 3-D virtual scene reconstruction process can be divided into four steps. The first one is the remote-sensing data preparation, where TM, SPOT5 and other remote sensing data were selected according to the characteristics of the cultural heritage of the Grand Canal and further subjected to data filtering and geometric correction. In the second step, the 3-D terrain modeling was carried out based on 3-D earth model segmentation and tile hierarchy system, where we fused and split remote sensing image and sampling spatial information for 3-D terrain model. The third step involves the modeling of local heritage with sophisticated modeling techniques sufficient to build a heritage 3-D model and to integrate terrain model with local scene through aerial orthophotos. Finally, in the fourth step the virtual scene integration is performed in a 3-D spherical system, where we designed a tree nodes system to assembly and manage multi-level and multi-type models of the Grand Canal. After these four steps are completed, the large-scale cultural heritage scene in 3-D spherical information system can be achieved.*

*Here, we address main challenges in virtual scene reconstruction of large-scale cultural heritage. This study can be valuable for regional and national cultural heritage protection as well as for Chinese government as a reference to infrastructural research, and finally for stimulation of other large-scale cultural heritage research around the world both in 3-D modeling and virtual scene organization.*

**KEYWORDS**

*Virtual scene construction, large-scale cultural heritage, the Grand Canal*

**1. INTRODUCTION**

Virtual reality technology has been applied in the protection of cultural heritage for about 20 years. Zahorik and Jenison [1] suggested that the couple of perception and action is crucial for determining the extent of presence, and mentioned a virtual scene may give the “presence” in ancient cultural heritage. Gaitatzes et al. and Miyazaki et al.[2, 3] presented an overview in modeling cultural heritage through observation. Their efforts were focused on three aspects: how to create geometric models of cultural heritage; how to create photometric models of cultural heritage; and how to integrate such virtual heritage components with real scenes. White et al. [4] proposed an architecture for integrating both the software and the hardware for digitization, management and presentation of virtual exhibitions, whereas Papagiannakis et al. [5] presented a case study of a real-time interactive digital narrative and real-time visualization of an ancient temple. In addition, Christou et al. [6] described the development and evaluation of a large-scale multimodal virtual reality simulation suitable for the visualization of cultural heritage sites and architectural planning. They referred “large-scale” to a haptic interface which was coupled with a realistic physics engine allowing users to experience and fully appreciate the effort involved in the construction of architectural components and their changes through the ages. According to Cabral et al.[7] X3-D is convenient in historic architectural reconstruction so that users might immerse themselves. Bruno et al. [8] summarized the complete methodology by a low-cost multimedia stereoscopic system for digital archaeological exhibition from digitization, management to user interfaces. Núñez Andrés [9] reviewed different techniques including massive capture techniques and traditional survey. They showed the advantages and disadvantages of each technique by applying them to the survey of the great Gate of Antioch.

However, the virtual scene constructions of cultural heritage sites in these studies are still restricted in local area or individual artifact. These methods or systems may create some intuitive 3-D model, but are not yet applicable to the processing of large-scale cultural heritage.

The primary characteristic of large-scale cultural heritage is its huge spatial span, such as Struve Geodetic Arc, the boundaries of the Roman Empire (including the Hadrian's Wall in UK, Der Obergermannisch-raetische Limes in Germany, etc. [10]), the Great Wall or Beijing-Hangzhou Grand Canal of China and so on. Spatial ranges of these large-scale cultural sites extend from hundreds to thousands kilometers.

A large spatial scale of cultural heritages is associated with difficulties in virtual scene construction. Although 3-D scanning and texture photography has been widely used on small objects modeling for many years, it might not be possible to acquire data of heritage spots over area of thousands of square kilometers especially with limited time and funding. Suppose the data are ready, just indexing and processing them could be a tremendous effort. Furthermore, management and rendering of the model output data would likely exceed the capability of ordinary computers. One possibility would be to adopt a Level Of Detail (LOD) concept to reduce data volume, but there are many 3-D model types such as tree, terrain, temples and the like which would hardly keep space consistency and could exceed a roam-able virtual scene.

How to construct a cultural heritage 3-D model for the large scale and how to manage and organize the scene have been so far difficult challenges not addressed in the literature.

We aimed at solving these two problems by taking the Grand Canal of China as a research subject, using remote sensing data sources for multi-scale modeling of cultural heritage, and adopting a 3-D geographic globe information system as a virtual scene organization platform. This paper explores and presents the overall framework and key steps in virtual scene construction of large-scale cultural heritage.

We first make a brief overview of the Grand Canal of China which is a typical example for a large-scale cultural heritage, and propose the overall roadmap of the virtual environment construction with the four crucial steps: (1) cultural heritage remote sensing data preparation; (2) terrain modeling; (3) local heritage modeling; and (4) virtual heritage scene integration in 3-D global system.

## 2. OVERVIEW OF THE BEIJING-HANGZHOU GRAND CANAL OF CHINA

The Beijing-Hangzhou Grand Canal is the world's oldest and longest canal, and also the largest in engineering scale. It is one of the "Two Great Works" of the ancient China (the other is the Great Wall). Beijing-Hangzhou Grand Canal belongs to a basin-wide cultural site. It connects five drainage basins including the Haihe River, Yellow River, Huaihe River, Yangtze River, and Qiantang River, with a total length of 1794 km, which is 16 times longer than the Suez Canal, and 33 times longer than the Panama Canal. Longer than the namely "the king of canal" the Main Turkmen Canal stretches over 400 km. In chronology, the Grand Canal is the earliest ancient canal which was dug up more than 2,500 years ago, and operated much earlier than the Panama Canal and the Suez Canal. Moreover, the Grand Canal runs through a huge number of ancient cities and other cultural heritages [11] with building age of these heritages from 770 BC to 1900, including 33 ancient cities and related pier, temples, pagodas, bridges, streets, factories, old kilns and ancient downtowns in 8 Chinese provinces.

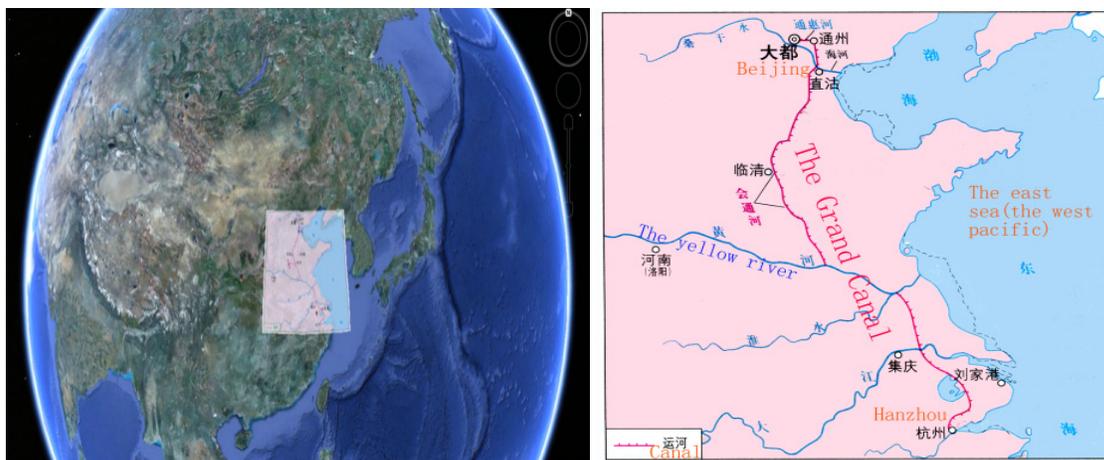


Figure 1. The vast area of the Beijing-Hangzhou Grand Canal

Beijing-Hangzhou Grand Canal is located in China's eastern flat and inhabitable area and its basin area accounts for 10% of China's land area, while Gross Domestic Product GDP accounts for 25% of China. The economic development, demand for industrial and residential land is

rapidly growing, but the environmental pollution is spreading as well. The protection of ancient ruins along the Grand Canal is facing increasing pressure, so the use of virtual reality technology to build virtual scene of the Beijing-Hangzhou Grand Canal is most likely the best means to help securing the Grand Canal Basin heritages, their history and culture.

On the other hand, the virtual scene construction of Beijing-Hangzhou Grand Canal is facing some unprecedented challenges. One of them is the large extent of the Beijing-Hangzhou Grand Canal, which makes it difficult to obtain 3-D data of thousands of kilometers. Another challenge is that the archaeological excavation on the Beijing-Hangzhou Grand Canal will constantly add to the new heritage modeling works, so it would be desirable (although not easy) to provide an open information platform and integrate these and subsequent models into the existing virtual scene.

### **3. METHODOLOGY**

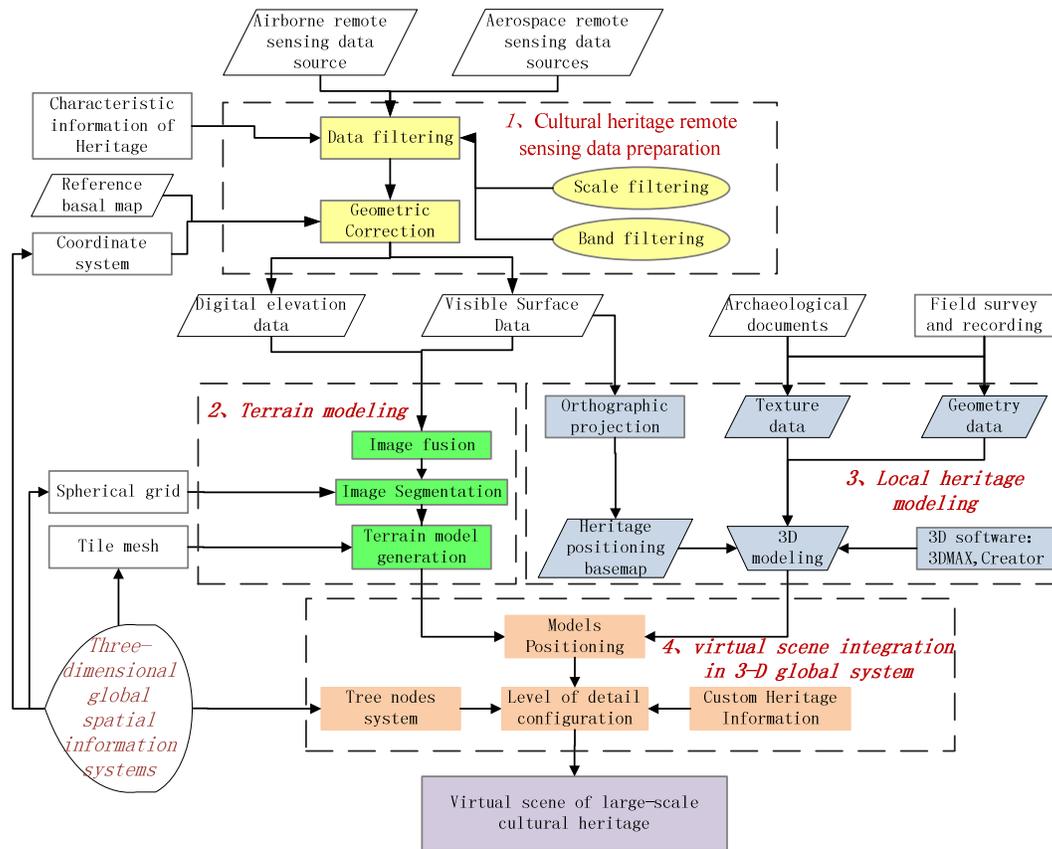
In order to solve the challenges in the virtual scene construction, we use remote sensing data sources for multi-scale modeling of cultural heritage, and adopt 3-D geographic globe information system as a virtual scene organization platform.

Remote sensing has been aiding the general approach of landscape archaeology since 1970s [12]. However, few studies introduced remote sensing data for 3-D cultural heritage modeling because of their low-accuracy. Here, we see a high potential of remote sensing in large-scale modeling of cultural heritage, particularly as basic data sources for environmental modeling.

It is widely known that GIS could provide an efficient way to integrate archaeological activities, data management, digital object representation and spatial analysis[13]. However, the ability to apply GIS techniques to achieve an integrated management of 3-D data and activities at a large scale is very limited in the field. Therefore, in the case of the Grand Canal and its subsequent 3-D modeling and system integration, we adopt a 3-D geo-spatial global information system, which is closer to digital earth than 3-D GIS, to establish an integrated environment for large-scale cultural heritage.

This virtual environment construction of large-scale cultural heritage based on remote sensing data sources and 3-D global geo-spatial information system can be divided into four parts in time sequence. These steps are: (1) remote sensing data preparation; (2) terrain modeling; (3) local cultural scene modeling; and (4) virtual heritage integration in 3-D global GIS. In the next sections, the contents and main issues of each step are analyzed in detail.

The overall research flow-diagram is shown in Figure 2 below:



### 3.1 Remote sensing data preparation for cultural heritage

Heritage remote sensing data preparation work is divided into two parts, data filtering (or data) and geometric precision correction.

Currently remote sensing has formed a multi-level, multi-angle, multi-field observation system from the ground to the air, and even space. From the 1960s onwards, technologies such as thermal infrared imaging, airborne synthetic aperture radar, multi-polar surface-penetrating radar and high-resolution space borne synthetic aperture radar have become sophisticated. Remote-sensing spectral bands from the earliest visible expand to near-infrared, shortwave infrared, thermal infrared, microwave direction, and these spectral extensions will adapt to a variety of data acquisition for material composition and geometrical shape. Synergies of large, medium and small satellites and combinations of high, medium, low orbits provide data which form a complementary series for specific domain with temporal resolutions ranging from a few hours to 18 days.

The remote sensing technology can nowadays provide more spatial information than ever before, but not all of the remote sensing data sources carry the characteristic data of spatial objects in desired location. Therefore, in the large-scale modeling study of cultural heritage, a proper selection is needed of the remote sensing data to express object characteristics of cultural heritage. Besides the simple judgment by spatial region and time stamp, the selecting work of remote sensing data can be divided into scale filtering and band filtering.

Scale is the key for understanding the complexity of spatial objects and is regarded as one of the standards of spatial information representation. Katsianis et al. [14] defined four kinds of scales associated with spatial phenomena, one of which is spatial resolution. An appropriate spatial resolution can reflect characteristics of the spatial structure of a specific target. Spatial resolution of remote sensing image represents the level of spatial detail and ability to separate spatial subject and its background, and also reflects the information hierarchy of the earth surface. For example, the riverway of the Grand Canal is 80 m wide on average, so 30 m resolution remote sensing data can be adequate for the riverway research and information extraction, but the same 30 meters remote sensing images are unable to express ancient buildings which are 10 meters wide on average. Therefore, in our study we selected the remote sensing image of optimal spatial resolution for expressing particular spatial objects inside the Grand Canal region according to their own characteristics (Table 1).

The Grand Canal cultural site is a complex spatial object with multi-scale structure, inside which objects have different spatial ranges. Thus, depending on the characteristic scale of the internal objects, we filtered remote sensing data sources at the first approach at virtual environment construction for the Grand Canal.

Table 1. The introduction of candidate remote sensing data

<b>Data type</b>	<b>introduction</b>	<b>average spatial resolution</b>	<b>sensitive spatial objects in the Grand Canal cultural site</b>
TM	U.S. Land Observing Satellite, three of seven observation bands are visible bands	30m	riverway of the ancient canal, lakes along the basin
SPOT5	French Earth Observation System (Système Probatoire d'Observation de la Terre)[15], provides 5 observations bands and panchromatic image	10m	change information of canal basin landscape (land types)
SRTM1	Shuttle Radar Topography Mission, provide terrain data from 60 degrees north latitude to 60 degrees south latitude with a total covering area of more than 119 million square kilometers	30m	the terrain topography along the Grand Canal
Airborne Remote Sensing	images are taken by airborne remote sensor, have higher spatial and temporal resolution, but narrower in spatial range comparing with satellite images Airborne photos	0.2m	contours of buildings and structures, roads and landmarks along the Grand Canal

Band filtering is a necessary step in the case of multi-spectral remote sensing data supplies, which requires selecting the best band or band sets to extract spatial objects and express spatial information effectively. Generally a remote sensor is only sensitive to specific spectral wavelength range which is named band. The original remote sensing images are all single band. In practice, according to three-color synthesis principle, the digital image processing systems place three different bands of remote sensing images on three channels which are assigned to red, green, blue to form a color image[16]. Specific details of band assembly are shown in Figure 3 .

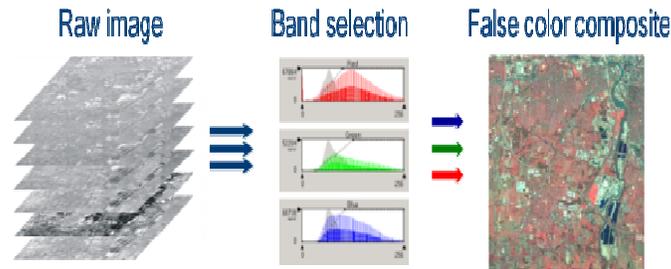


Figure 1. Band assembly of Spot5 data for The Grand Canal

There are three main principles for the best band selection:

(1) The largest amount of information it carries; (2) the smallest possible correlation among bands; (3) the largest difference among spatial object types which the band is sensitive to. The best band combination is three bands with abundant information content, the smallest correlation and distinguishable sensitive spatial objects [17]. However, despite each band's own characteristics, there are more or less information duplication and redundancy among hyperspectral remote sensing data. Consequently, the selection should be based on the spectral characteristics of observation targets. For example, TM1,4,5 would be the best band combination for settlements and waters interpretation[18], while TM3,4,5 would be the best band combination for farmland, woodland, grassland interpretation[19]. When using remote sensing data to construct a 3-D scene of cultural heritage, selection of band combinations should be based on the characteristics of spatial objects inside the heritage site.

The core heritage trail and main observation target of the Grand Canal is the riverway, of which the main body is water. The spectral characteristic of water depends on its material composition, but also reflects the various water states. Natural water bodies have significantly higher absorption for 0.4 ~ 2.5  $\mu\text{m}$  electromagnetic wave than most other surface features [20]. In the infrared, water absorbs more energy than in visible light, while vegetation, soil in the two bands have smaller energy absorption, and a higher reflection, that results in significant difference between the water and the vegetation or soil in the two bands. Although the panchromatic band is not more sensitive to the Grand Canal water body than other spatial objects, the panchromatic band provides a high spatial resolution of 2.5 m, which has a good effect for the spatial information expression of the overall Grand Canal region. For these reasons, we chose the near infrared (XS3), mid infrared (B4), and panchromatic band (PAN) for remote sensing image synthesis of the Grand Canal heritage (Table 2).

The purpose of the geometric correction is to correct errors caused by non-systemic factors, and is based on mathematical models of geometric calibration. This correction is also a process which combines translation and rotation to project image onto a plane through homonymy points from remote sensing images and underlay reference map to place the spatial objects right on the corresponding reference position after correction [21].

In the Grand Canal remote data preparation, we used the method of least squares to calculate imagery geometric correction, in which  $(t+1) \times (t+2) / 2$  equals to 6 control points similarly to as described elsewhere [22]. In order to promote accuracy, in each image we select at least 15 control points for image registration. Generally there are two principles in control point selection for large-scale cultural heritage:

(1) The control points should be evenly distributed over the entire heritage site.

- (2) The control points should optimally be the corner points of a permanent feature rather than of any removable points, such as docks or road crossings in the case of the Grand Canal.

Table 1. The bands features of SPOT5 sensor

Spectral band	wavelength range ( $\mu\text{m}$ )	ground resolution (m)	The main application areas
XS1:Green	0.50 ~ 0.59	10	This band has part transmission in water body and great reflectance in healthy green plants, it can distinguish vegetation types and assess crop growth.
XS2:Red	0.61 ~ 0.68	10	This band provides measurable plants chlorophyll absorption rate, and in turn plant classification, can distinguish artificial feature types too.
XS3: near infrared	0.78 ~ 0.89	10	This band is greatly absorbed in water and is used to draw water body boundaries, detect the content of aquatic organisms.
B4: mid infrared	1.58 ~ 1.75	20	Usually vegetation, water, soil have obvious gray level difference in this band, so this band is used to detect vegetation, water content and soil moisture, and distinguish the difference between clouds and snow.
PAN	0.48 ~ 0.71	2.5	With the highest spatial resolution in SPOT5, this band can be used for forestry research and planning, urban planning and large scale thematic mapping.

### 3.2 Terrain modeling of heritage site

Heritage site terrain modeling refers to processing of remote sensing data of large-scale cultural heritage into a 3-D terrain model so that information systems can load and render as background-model of the virtual scene. It is an important step of the large-scale cultural heritage modeling. On the one hand, numerous reports [23, 24] suggested that the terrain features are closely related with formation and development of large-scale cultural heritage areas. On the other, with 3-D terrain models, viewers could analyze and rebuild the ancient ruins more intuitively, and present more reasonable explanation of different cultural phenomena from archaeological surveys and excavations than without them [25].

The terrain modeling study is based on 3-D global information systems. The main reason for the system adoption is no limitation on the scale of possible terrain model which may cover a huge spatial range.

Terrain modeling includes image fusion, image segmentation and terrain model generation three steps, and how to perform these steps is determined by 3-D Geographic Information System 3-D Earth model structure. Therefore, before terrain modeling work, an introduction to 3-D Earth model is necessary (Figure 4 ).

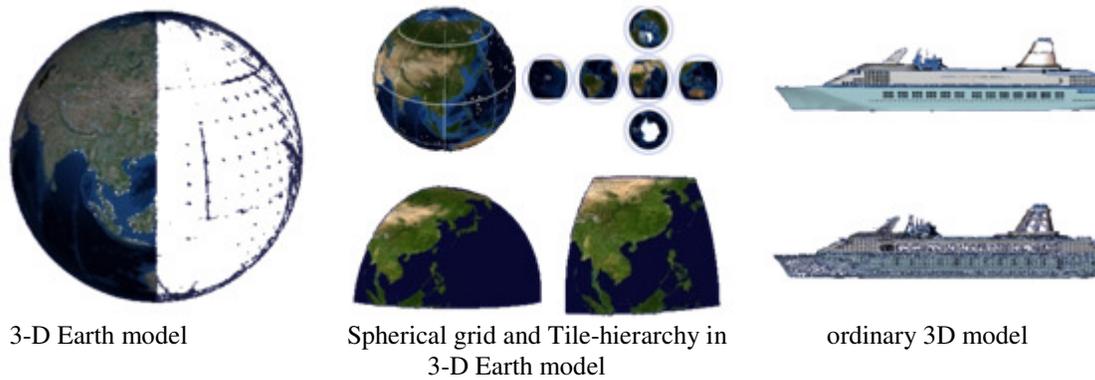


Figure 2. The comparison between 3-D Earth model and ordinary 3D model

3-D Earth model, which is a combination of multiple parts, differs from an ordinary 3-D model. The division rules can be classified into two categories: spherical grid and tile hierarchy. Purpose of the both is to reduce the model data throughput during system operation, but they are not the same in principles and functions.

Spherical grid is a seamless mesh system of the Earth ellipsoid surface with the classic subdivisions called graticules. Cells in spherical grid could be uniform or non-uniform (irregular). Clearly, the morphology of uniform spherical grid cell is more stable with regular borders, and is more convenient to prepare and assembly 3-D data in the 3-D Earth model than that with the irregular borders.

Tile-hierarchy is a hierarchically organized system of spatial data with each tile in the hierarchy being a specific 3-D model. Basic idea of tile-hierarchy is to subdivide a parent tile into a number of child tiles under same data volume constraint, the smaller range a tile covers, and at the higher resolution it could possess [26]. Owing to this subdivision, when spatial data are required, an appropriate tile with the closest scale is delivered so that to reduce the access time and data load during User-System interaction.

As a concrete 3-D model, 3-D Earth tile is not only the direct target for spatial data integration, but also the environmental information carrier in the case of the Grand Canal. We were able to seamlessly integrate spatial data or 3-D models into the tile.

Tile in 3-D Earth model, similarly to other 3-D visualization models, is composed of two parts: geometry mesh and texture. The mesh information and texture information in tile are derived from existing spatial data. Tile mesh is used to represent geometrical shape of the earth surface with each vertices in the mesh expressed in spherical coordinates  $(r, \theta, \varphi)$ . Geometric resolution of tile means separation of the zenith angle  $(\theta)$ , azimuth angle  $(\varphi)$  between tile vertices. The radial distance  $(r)$  of tiles vertex in earth science field usually refers to elevation value which is derived from remote sensing data products such as elevation SRTM in this case. The tile texture is used to indicate shape and color of earth surface, in which case the tile texture is the visible spectrum data from the hyper spectral remote sensing image such as SPOT5 or TM.

Given the basic platform and the structure of the 3-D earth model, the 3-D terrain modeling process is to assign spatial information from remote sensing data sources into tile vertex.

Firstly, image fusion here was used to integrate different spatial data source into one tile vertex with predefined weights. In the case of the Grand Canal modeling, we based on the same UTM

coordinate system fuse TM remote sensing images; the weight of overlapping portion is the arithmetic mean, as shown in Figure 5:

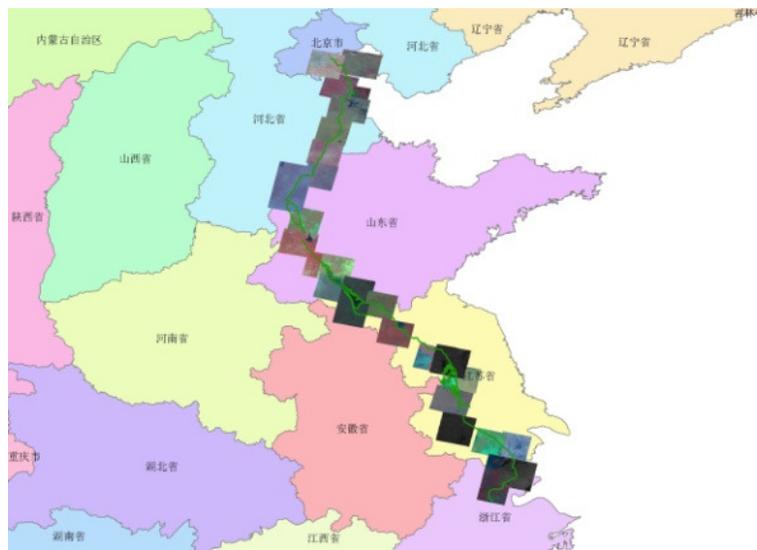


Figure 3. The Spot5 and TM data fusion along the Grand Canal

Secondly, remote sensing image segmentation must be consistent with the borders definition of spherical grid and tile hierarchy, as otherwise it may cause terrain dislocation in the entire virtual scene. In the case of the Grand Canal modeling, in accordance with the tile with 10 meter spatial resolution and 30 km side-length, we cut TM5 images and elevation images as demonstrated in Figure 6:

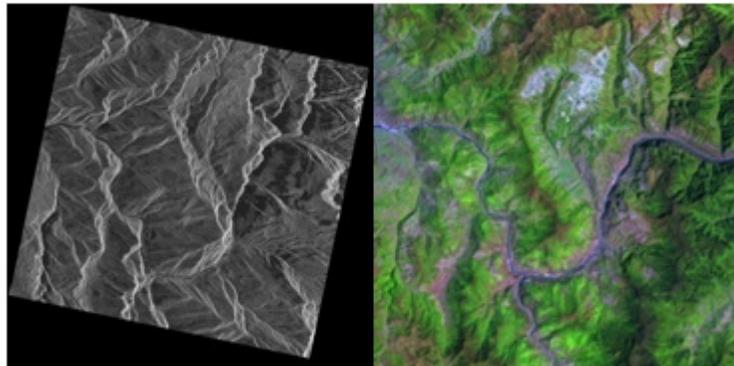


Figure 4. Segmentation of TM5 image and elevation image(rotated) according to the tile boundary

Thirdly, depending on predefined tile meshes of each scale, we sampled remote sensing images to every tile vertex. The generation process of a specific 3-D terrain model of a heritage site was as presented in Figure 7 and linear interpolation of vertices to fill the final model was as shown in Figure 8.

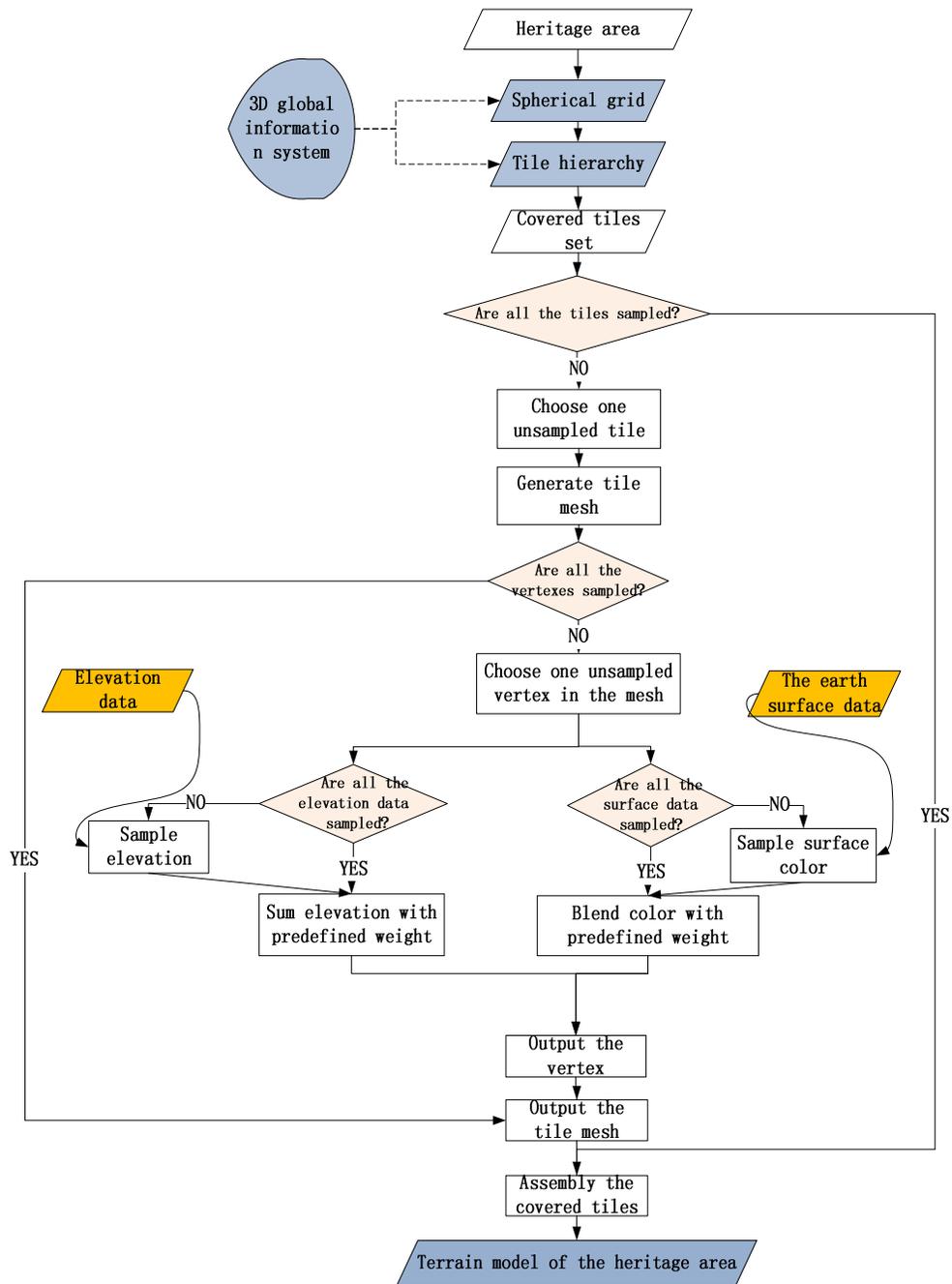


Figure 5. Generation process of a specific 3-D terrain model



Each vertex of tile mesh has sampled color and elevation from remote sensing data



Linear interpolation of vertexes to fill the final model

Figure 6. 3-D terrain modeling based on previous image segmentation

### 3.3 Local heritage modeling

Local heritage refers to centralized heritage places in the large-scale cultural heritage area. Compared with terrain modeling, local heritage modeling has two differences:

One difference is the higher spatial resolution of a model. A 3-D terrain model is under the resolution limitations of remote sensing data that the maximum is not more than 0.2 meters, while in local heritage modeling where we can use a laser radar or a high-precision camera for data acquisition, the spatial resolution can reach even centimeter or millimeter resolution level;

The other difference is a richer type of modeling objects in local heritage. Although a 3-D terrain model covers a wide range, its main purpose is mostly to provide a natural background model which brings basic information on topography and landforms. While the local heritage modeling includes a variety of static spatial entities in cultural heritage area, specific modeling objects includes buildings, vegetation, bridges, piers, heritage and other artificial or natural objects.

Local heritage modeling uses mainly a 3-D model of small-scale cultural heritage. So far its data collection and modeling process, data integration and key technical problems have been thoroughly discussed in many previous studies. For example, Akca [27] used a structured light system for data acquisition, Carmel et al.[28] classified entities of cultural heritage for targeted digitalization, and Alsadik et al.[29] presented a camera network for image-based modeling of cultural heritage. Hug and Gonzalez-Perez [30] qualitatively evaluated three modeling techniques derived from information system engineering to represent cultural heritage domain concepts. More detailed technology of local heritage modeling has been described elsewhere [2, 6, 23, 31, 32].

Both local heritage models and 3-D terrain models are parts of the large-scale virtual scene. After local heritage modeling, different models in geographic global information system need to be integrated. This integration is solved by using the orthographic projection (Figure 9).

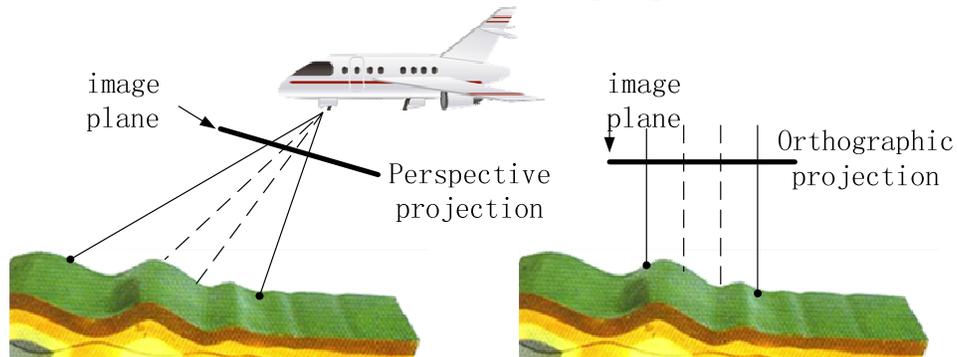


Figure 7. The necessity of orthographic projection in airborne remote sensing image

The projection of an airborne remote sensing image is the central projection which is associated with distortion. Because photosensitive surface tilts and undulating ground generate point displacements in an image, raw airborne images cannot objectively and accurately indicate the shape and location of spatial features. The images have to be processed in order to obtain orthophotos, which are the photographs orthographically projected, and the process is called ortho-correction [33].

Orthophotos, in particular airborne orthophotos are important basal data of cultural heritage 3-D modeling. They have high spatial resolution, allowing for clear identification of buildings, and silhouette of landmarks. As airborne orthophotos are perpendicular to the ground, they reflect the true ground position and topological relations, which can provide spatial orientation and positioning information for high-precision 3-D model to guarantee the accuracy of 3-D spatial measurement and analysis. The import to the 3-D global information system is presented in Figure 10.



Figure 8. Airborne remote sensing image import into 3-D global information system

As shown in Figure 11, we obtained an airborne remote sensing image of the ancient canal sluice ruins after ortho correction. This led to elimination of the shadows and each pixel on the image regaining its right place within geographic coordinates. Thus, not only the corrected image can be imported into a 3-D modeling software such as Basal Map, 3-Dmax or Creator, but it can also be

used as a texture layer, for the terrain model to sample from. After integration in 3-D global information systems, the identical geographic coordinates provided by orthophoto become the key bridge between 3-D earth model (terrain model) and the heritage virtual scene, so the high precision of orthophoto ensures accuracy and consistency of the integration.



Figure 9. Airborne remote sensing image of a sluice on the Grand Canal before and after orthographic projection

Orthophoto production often requires special equipment. In the case of central projection aerial photographs of flat ground using a mechanical optical instrument could be sufficient. On an undulating ground or a bumpy flight, orthographic projection device such as an optical-mechanical differential rectification instrument and computer numerically controlled analytic projector is available. In the case of dynamic remote sensing which have not captured predefined parameters, all-digital correction machine is usually used to produce the orthophoto[34].

As shown in Figure 12, we located the specific local heritage model based on airborne orthophoto in the 3-DMax software.

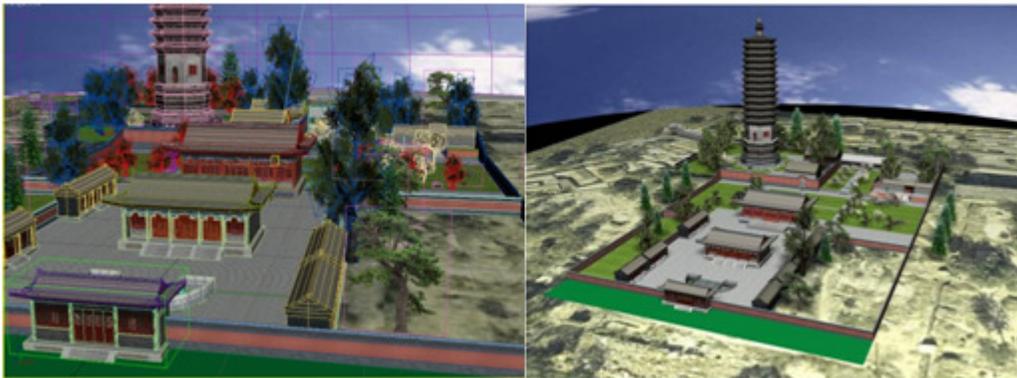


Figure 10. Local heritage modeling and positioning based on Orthophoto in 3-D Max

### 3.4 Virtual scene integration in 3-D spherical system

In the final integration step, in order to embed heritage models into 3-D global information systems seamlessly, we have established an integrated tree of 3-D models (Figure 13).

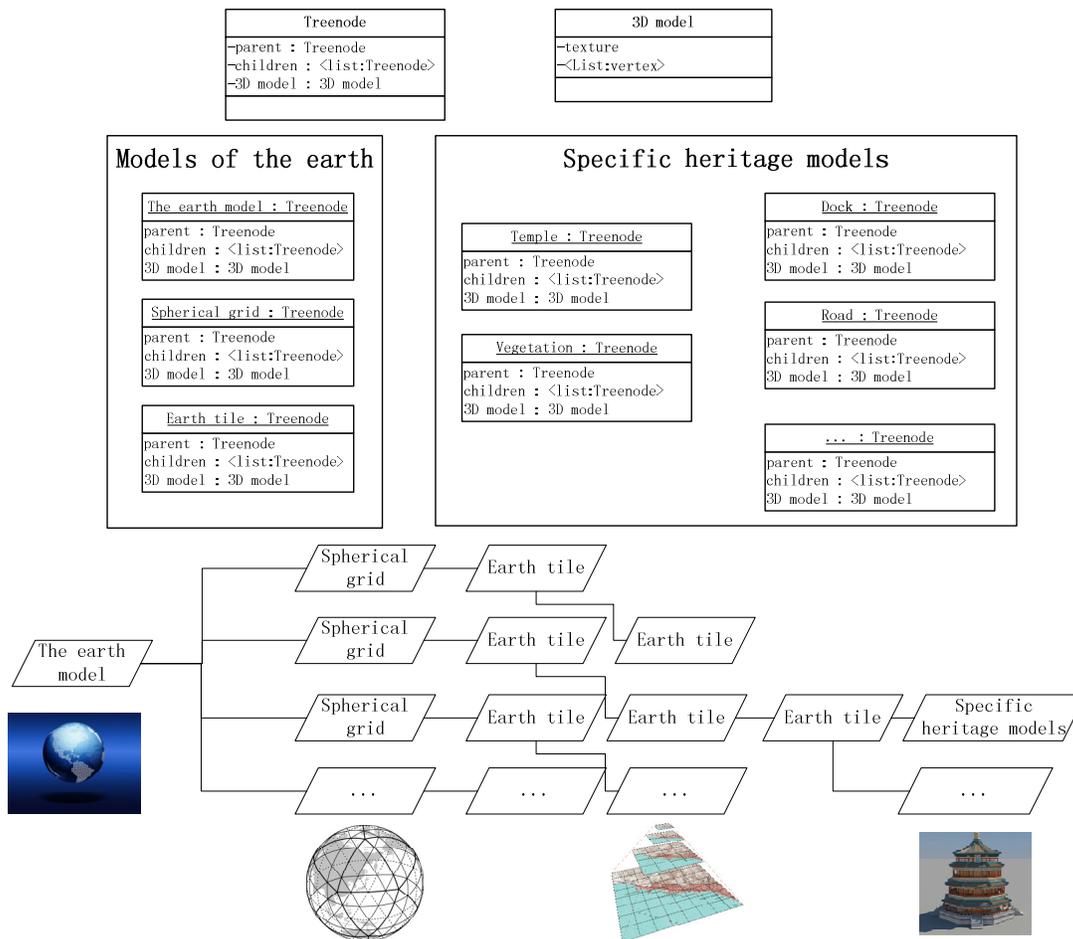


Figure 11. The integrated tree system for 3-D virtual heritage scene organization in 3-D global information system

A tree node object encapsulates each 3-D model, and it may have no more than 1 parent node and unlimited child nodes. This treelike organization for large-scale virtual scene has three characteristics:

- (1) All levels of spatial objects are under identical geographic coordinate system
- (2) Spatial ranges of all child nodes do not exceed the parent tree node.
- (3) In LOD configuration, models-switch from the parent node to child node by viewpoint movement is seamless. It means there is no place to see both the parent and its child nodes and there is also no place to see nothing at heritage site.

This treelike organization has two advantages for large-scale heritage scene:

- (1) The data object and user interface definitions in heritage scene are uniform. Every 3-D model of mutli-scale or different types are all tree nodes, such as terrain, roads, buildings, lights, controllers, triggers, particle systems, etc. It helps achieving flexibility in virtual scene assembly so that each node can be assigned to one another as parent or child. Through a consistent interface, operations like node add, delete, property editing and others are very convenient.

(2) Tree level reflects spatial cover. This implicit spatial description makes it simple and easy to understand spatial relations among 3-D models. Through the view of nodes' bounding box, we can intuitively choose operations including add, delete, edit for virtual scene assembly. Furthermore, each tree node could have its custom properties which are stored in attribute list that promotes the extensibility of the virtual scene to present extra information about history and culture.

## 4. RESULTS

The results from the present study can be divided into two aspects. One was the establishment of a presentation platform for integrated large-scale natural and cultural heritage. The other was to assemble the terrain models of the entire Grand Canal with local heritage models of Jining(ancient city), Tongzhou(ancient city), and the Dragon King Temple which located at the junction between the Grand Canal and Wenshui river. The average distance between these three local heritages was 300 km .

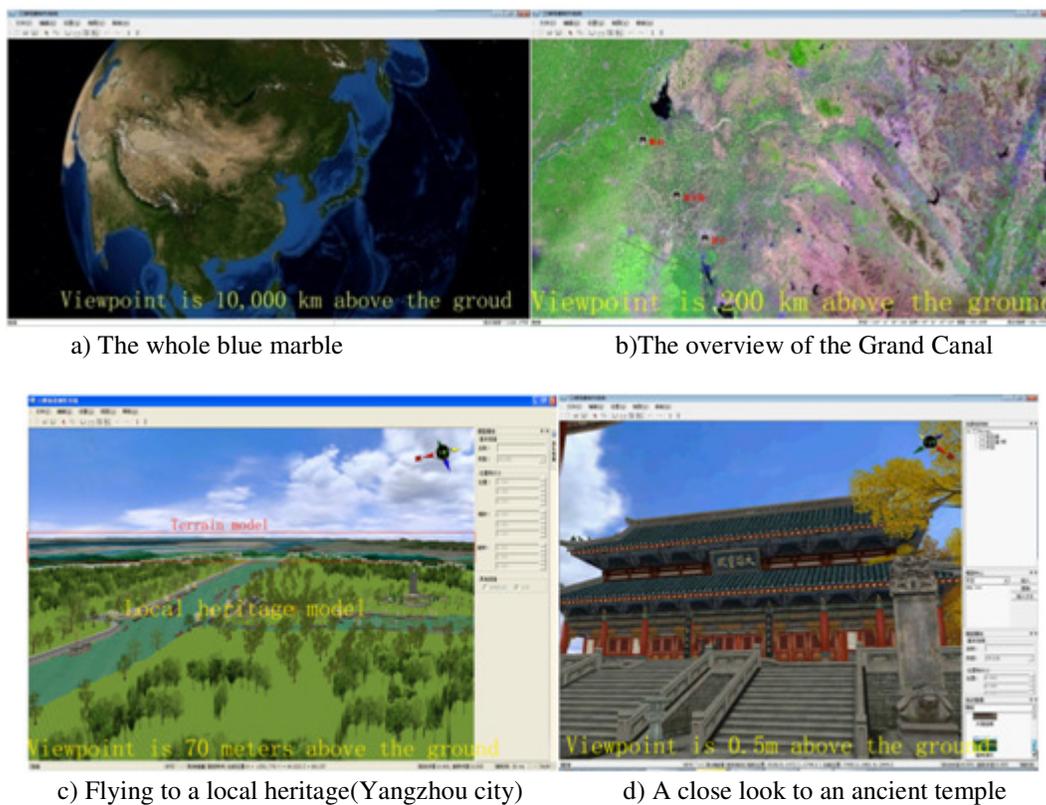


Figure 14 The integrated scene of the Grand Canal in the 3-D global information system

## 5. DISCUSSION

In order to establish three-dimensional model of large-scale cultural heritage for the present study data selection, processing, combination with a 3D Earth model, integrated management of 3D GIS and a comprehensive workflow were completed.

One novel thing about the study is that the remote-sensing data source was introduced. In the modeling of cultural heritage this type of data source is not typically used. The reason is that it is collected from too far away, and the accuracy of the data cannot reach the standard for 3D modeling. In addition, the shooting angles almost do not reflect the information on the sides of cultural heritage at all[35]. However, here we found advantages of remote sensing data source in modeling of large-scale cultural heritage.

For example, remote-sensing data source can provide spatial information that reflects various types of surface features in the cultural heritage. The reason is that the existing remote sensors cover a rather wide wavelength range, and for any surface feature, corresponding sensitive data can be found.

Remote sensing data sources can help to achieve quickly 3D modeling of large-scale terrains and landforms. This is particularly important in the modeling of large-scale cultural heritage that crosses hundreds of kilometers.

Therefore, based on the electromagnetic wave reflection characteristics of cultural heritage, after careful data screening and matching, remote sensing data can easily provide the information needed for modeling of cultural heritage.

Another development of the present study was that the 3-D global GIS was used for integration of scenes of cultural heritage at a large scale although this strategy was originally inspired by Google earth experience. In the 3D GIS, the space that can be explored is unlimited. In addition, differences in the spatial scales of terrains and surface details can be directly viewed.

When the present project was launched in 2007, Rome Reborn in Google Earth (RR project)[[http://romereborn.frischerconsulting.com/project\\_news.php](http://romereborn.frischerconsulting.com/project_news.php)] was not yet on line. Although our ultimate interface looks very similar to that of Google Earth, there are certain differences. Firstly, in the RR project[36], established models are released to Google earth as kml layer file, the information on cultural heritage can only be linked through the web, and visualization of the information is prioritized. In our project, for the scenes of cultural heritage a more specialized professional structure was adopted. The cultural heritage model and the 3D Earth model were integrated using a tree structure. In addition, at each node aside from the 3D visual information, custom culture, heritage information can be added. In this way, management and retrieval of cultural heritage information can be further developed. Secondly, the paper [36] published from the RR project did not disclose the method used to prepare spatial data and process of integrating the Earth model and the cultural heritage model. This may be because as two different groups were responsible for modeling the Earth and modeling the ancient buildings, the information available to the author was incomplete.

Although our modeling developments may have still some potential for improvement, we find our study a robust and accurate representation which works extremely well for the spatial features of the Grand Canal. All parts of our final result focus on the Grand Canal heritage, allowing users to better understand the entire picture of the Grand Canal heritage. While the 3D model integrated in the RR project was focused on the scope of the ancient city of Rome, which is a relatively small-scale cultural heritage, in our project, processing of remote sensing data terrain modeling over thousands of kilometers was completed and three local heritage models with 300 km space between were also established.

Due to limited funding and staff, currently we have only over 200 high-precision three-dimensional models. Nonetheless, our study provides a basic framework, and an import interface for new models. This will provide continuous support for the presentation and protection of the Grand Canal heritage in the future.

As the present study focused on establishing large-scale modeling and integration framework for cultural heritage in space, it can be expected there will be many follow-up studies. The construction of large-scale cultural heritages often takes decades or even centuries. In addition to the spatial modeling and integration framework as explored in the present study, it is also of great significance for archaeological research and cultural education to study the modeling method and integration framework of large-scale cultural heritages in time. Another point our study makes is that the way of exploring scenes in large-scale cultural heritages should be different from that in small-scale heritages. How to allow the users to better perceive multi-scale spatial characteristics of scenes in large-scale cultural heritages may also be worth investigating.

## 6. CONCLUSION

Large-scale cultural heritage is a great concentration of national or regional history and culture with high research and conservation value. The current studies of the 3-D cultural heritage scene reconstruction mainly focused on local scale heritage due to a lack of modeling methods and scene organization for large-scale cultural heritage. We took the Grand Canal of China as an example for systematic reconstruction at a larger scale. We introduced a variety of remote sensing data sources for large-scale cultural heritage modeling, and then adopted a 3-D geographic global information system for large-scale 3-D scene organization and management. A four-step 3D virtual reconstruction was developed and successfully applied.

This study presented methods and key technology in aspects of data sources and model organization to solve reconstruction problem of large-scale cultural heritage with high efficiency and accuracy.

In data source aspects, the large regional remote sensing data and their multi-band images seem suitable for 3-D reconstruction of large-scale cultural heritage. In particular, existing various data sources were more than sufficient in providing the basal data for cultural heritage terrain modeling.

The 3-D spherical information system seems also suitable for organization and integration of large-scale 3-D scene.

The identical coordinate system was able to keep uniform all of the 3-D models in the virtual scene and provided integrated spatial analysis functionalities like area, round, distance calculation.

The tree level showed spatial relationship between the 3-D models. Parent-child nodes indicated include-spatial-relationship. The spatial extent of all child nodes did not exceed that of the parent's, or sibling nodes which are adjacent to each other, and thus it provided a basis for spatial relational query.

The root node of the platform is the Earth model ensured cultural heritage virtual scene could be integrated without a spatial limit. This platform could import both the 3-D terrain model and the local heritage model of every large-scale cultural heritage to spherical grid and tile-hierarchy of the Earth model, and render them all together, which cannot be done in the other 3-D system.

Our study aimed at addressing the main challenges in virtual scene reconstruction of large-scale cultural heritage by application of remote sensing data and 3-D global GIS from spatial information field. This proved the efficiency and capability on the case study of the Grand Canal.

We believe this study can be valuable for regional and national cultural heritage protection, for Chinese government as an infrastructural research, and as a good reference for other large-scale cultural heritage around the world both in 3-D modeling and virtual scene organization.

## REFERENCES

- [1] Zahorik, P. and R.L. Jenison, Presence as being-in-the-world. *Presence: Teleoperators and virtual environments*, 1998. 7(1): p. 78-89.
- [2] Gaitatzes, A., D. Christopoulos and M. Roussou. Reviving the past: cultural heritage meets virtual reality. in *Proceedings of the 2001 conference on Virtual reality, archeology, and cultural heritage*. 2001: ACM.
- [3] Miyazaki, D., et al., The great buddha project: Modeling cultural heritage through observation, in *Modeling from reality*. 2002, Springer. p. 181-193.
- [4] White, M., et al. ARCO-an architecture for digitization, management and presentation of virtual exhibitions. in *Computer Graphics International*, 2004. *Proceedings*. 2004: IEEE.
- [5] Papagiannakis, G., A. Foni and N. Magnenat-Thalmann. Real-Time recreated ceremonies in VR restituted cultural heritage sites. in *CIPA 19th International Symposium*. 2003.
- [6] Christou, C., et al. A versatile large-scale multimodal VR system for cultural heritage visualization. in *Proceedings of the ACM symposium on Virtual reality software and technology*. 2006: ACM.
- [7] Cabral, M., et al. An experience using X3D for virtual cultural heritage. in *Proceedings of the twelfth international conference on 3D web technology*. 2007: ACM.
- [8] Bruno, F., et al., From 3D reconstruction to virtual reality: A complete methodology for digital archaeological exhibition. *Journal of Cultural Heritage*, 2010. 11(1): p. 42-49.
- [9] Núñez Andrés, A., et al., Generation of virtual models of cultural heritage. *Journal of Cultural Heritage*, 2012. 13(1): p. 103-106.
- [10] Van Gorp, B. and H. Renes, A European cultural identity? Heritage and shared histories in the European Union. *Tijdschrift voor economische en sociale geografie*, 2007. 98(3): p. 407-415.
- [11] Qiao-yi, C., The Grand Canal—On the Protection of Canal Culture. *Journal of Hangzhou Teachers College*, 2005. 3: p. 000.
- [12] Clark, C.D., S.M. Garrod and M.P. Pearson, Landscape archaeology and remote sensing in southern Madagascar. *International Journal of Remote Sensing*, 1998. 19(8): p. 1461-1477.
- [13] Katsianis, M., et al., A 3D digital workflow for archaeological intra-site research using GIS. *Journal of Archaeological Science*, 2008. 35(3): p. 655-667.
- [14] Cao, C. and N.S. Lam, Understanding the scale and resolution effects in remote sensing and GIS. *Scale in remote sensing and GIS*, 1997. 57: p. 72.
- [15] Chevrel, M., M. Courtois and G. Weill, The SPOT satellite remote sensing mission. *Photogrammetric Engineering and Remote Sensing*, 1981. 47: p. 1163-1171.
- [16] Vrabel, J., Multispectral imagery band sharpening study. *Photogrammetric Engineering and Remote Sensing*, 1996. 62(9): p. 1075-1084.
- [17] Schowengerdt, R.A., *Remote sensing: models and methods for image processing*. 2006: Academic press.
- [18] Markham, B.L. and J.L. Barker, Spectral characterization of the Landsat Thematic Mapper sensors. *International Journal of Remote Sensing*, 1985. 6(5): p. 697-716.

- [19] Yang, C., J.H. Everitt and D. Murden, Evaluating high resolution SPOT 5 satellite imagery for crop identification. *Computers and Electronics in Agriculture*, 2011. 75(2): p. 347-354.
- [20] DENG, J., et al., An Effective Way for Automatically Extracting Water Body Information from SPOT-5 Images. *Journal of Shanghai Jiaotong University (Agricultural Science)*, 2005. 2: p. 198-201.
- [21] Gonçalves, H., J.A. Gonçalves and L. Corte-Real, Measures for an objective evaluation of the geometric correction process quality. *Geoscience and Remote Sensing Letters, IEEE*, 2009. 6(2): p. 292-296.
- [22] Richards, J.A., *Remote sensing digital image analysis: an introduction*. 2013: Springer.
- [23] Guarnieri, A., F. Remondino and A. Vettore, Digital photogrammetry and TLS data fusion applied to Cultural Heritage 3D modeling. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2006. 36(5).
- [24] Doneus, M. and Briese, C., Digital terrain modelling for archaeological interpretation within forested areas using full-waveform laserscanning. In: M. Ioannides, D. Arnold, F. Niccolucci and K. Mania (Editors), *The 7 International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2006)*
- [25] Gruen, A., Reality-based generation of virtual environments for digital earth. *International Journal of Digital Earth*, 2008. 1(1): p. 88-106.
- [26] Gonzalez-Escribano, A., et al., An extensible system for multilevel automatic data partition and mapping. *IEEE Transactions on Parallel and Distributed Systems*, 2013. 99(1): p. 1.
- [27] Akca, D., 3D Modeling of cultural heritage objects with a structured light system. *Mediterranean Archaeology and Archaeometry*, 2012. 12(1): p. 139-152.
- [28] Carmel, D., N. Zwerdling and S. Yogev. Entity oriented search and exploration for cultural heritage collections: the EU cultura project. in *Proceedings of the 21st international conference companion on World Wide Web*. 2012: ACM.
- [29] Alsadik, B., M. Gerke and G. Vosselman. Optimal Camera Network Design for 3D Modeling of Cultural Heritage. in the *Proceedings of the XXII ISPRS Congress*, Melbourne, Australia. 2012.
- [30] Hug, C. and C. Gonzalez-Perez, Qualitative evaluation of cultural heritage information modeling techniques. *Journal on Computing and Cultural Heritage (JOCCH)*, 2012. 5(2): p. 8.
- [31] Papagiannakis, G., A. Foni and N. Magnenat-Thalmann. Real-Time recreated ceremonies in VR restituted cultural heritage sites. in *CIPA 19th International Symposium*. 2003.
- [32] Pavlidis, G., et al., Methods for 3D digitization of cultural heritage. *Journal of Cultural Heritage*, 2007. 8(1): p. 93-98.
- [33] Shimada, M., Ortho-rectification and slope correction of SAR data using DEM and its accuracy evaluation. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 2010. 3(4): p. 657-671.
- [34] Du, Q., et al. Digital Orthoimage Generation with Low Altitude Photogrammetric System Based on Unmanned Airship. in *Image and Data Fusion (ISIDF), 2011 International Symposium on*. 2011: IEEE.
- [35] Pu, S. and G. Vosselman, Knowledge based reconstruction of building models from terrestrial laser scanning data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2009. 64(6): p. 575-584.
- [36] Wells, S., et al. Rome Reborn in Google Earth. in *Making history interactive, 37th proceedings of the CAA conference*. 2009.

## AUTHORS

### Dr. Jian Tan

Ph. D. in WEB Geographic information science

Affiliation: Key Laboratory of Digital Earth, Center for Earth Observation and Digital Earth, CAS (Visualization Technology department, International Centre on Space Technologies for Natural and Cultural Heritage under the auspices of UNESCO)

Research Interests: 3D Digitalization Aiding Cultural Heritage Archaeology Or Heritage Conservation; Digital Earth and Its Applications



# A PREFIXED-ITEMSET-BASED IMPROVEMENT FOR APRIORI ALGORITHM

Yu Shoujian<sup>1</sup>, Zhou Yiyang<sup>2</sup>

College of computer science and technology,  
Donghua University, Shanghai, 201600, China

<sup>1</sup>jackyysj@dhu.edu.cn

<sup>2</sup>yiyang0203@foxmail.com

## **ABSTRACT**

*Association rules is a very important part of data mining. It is used to find the interesting patterns from transaction databases. Apriori algorithm is one of the most classical algorithms of association rules, but it has the bottleneck in efficiency. In this article, we proposed a prefixed-itemset-based data structure for candidate itemset generation, with the help of the structure we managed to improve the efficiency of the classical Apriori algorithm.*

## **KEYWORDS**

*Data mining, association rules, Apriori algorithm, prefixed-itemset, hash map*

## **1. INTRODUCTION**

With the rapid development of computer technology in various sectors, the data generated by different industries are becoming more and more, but how to get valuable information from the big data has become a new problem. Data mining, that is data knowledge discovery, came into being in this backdrop. Data mining is to excavate the implied, unknown, interesting knowledge and rules from a large number of data <sup>[1]</sup>. Association rules is an important part of data mining, it was first put forward by R.Agrawal, mainly to solve the customer transaction association rules between sets of items in the transaction library <sup>[2]</sup>. In the following year, R.Agrawal proposed the most classical algorithm to calculate association rules, that is Apriori algorithm <sup>[3]</sup>, which is to infer the (k+1) – itemsets by the k- itemsets.

However, due to the computing bottleneck of Apriori algorithm when calculating the candidate set, in recent years there have been many improved algorithms of the traditional Apriori algorithm from different aspects. Chun-Sheng Z proposed an improved Apriori algorithm based on classification <sup>[4]</sup>. Jia Y improves the algorithm from the aspect of transaction database partitioning and dynamic itemset planning <sup>[5]</sup>. Shuangyue L proposed an improved algorithm based on the matrix of database to enhance the efficiency of calculating <sup>[6]</sup>. Wang P proposed an optimization method to reduce the search times of the transaction library to improve the efficiency <sup>[7]</sup>. Vaithiyathan V uses the method of compressing the transactions of the similar interests in the database to improve the efficiency of the algorithm <sup>[8]</sup>. Lin X implements Apriori

algorithm based on Map Reduce to improve the candidate sets of large amounts of data generation efficiency<sup>[9]</sup>. Zhang first analyze the characteristic of the data, that is medical data, and then combine the characteristics of the data to improved Apriori algorithm<sup>[10]</sup>. Wu Huan proposed an improved algorithm IAA, which adopts a new count-based method to prune candidate itemsets and uses generation record to reduce total data scan amount<sup>[11]</sup>. Wang Yuan proposes an improved item constrain association rules mining algorithm, which improves traditional algorithm in two aspects: trimming frequent itemsets and calculating candidate itemsets<sup>[12]</sup>. Lin Ming-Yen proposes three algorithms, named SPC, FPC, and DPC, to investigate effective implementations of the Apriori algorithm in the MapReduce framework<sup>[13]</sup>. Chai Sheng proposes a novel algorithm so called Reduced Apriori Algorithm with Tag (RAAT), which reduces one redundant pruning operations of C2<sup>[14]</sup>.

This article will be focus on the two concrete steps of classical Apriori algorithm, namely connecting step and the pruning step, using a new prefix-itemset-based storage, combining the fast lookup feature of hash tables to improve the efficiency. This paper will first describe the classical Apriori algorithm and its shortcomings, then specifically describe the improvements, and finally introduce the comparisons of efficiency of classical Apriori algorithm and improve Apriori algorithm on specific data sets.

## 2. APRIORI ALGORITHM

### 2.1. Apriori algorithm introduction

Apriori algorithm is a classical algorithm for frequent itemset mining association rules, the basic idea of the algorithm is to use an iterative approach layer by layer to find the frequent. The algorithm will first obtain k-itemsets, and then use the k- itemsets to explore (k+1)-itemsets. First, let's introduce the priori knowledge of frequent itemsets, which is, any subset of a frequent itemset is also a frequent itemset. Apriori algorithm uses the prior knowledge of frequent itemsets, first to find the collection of frequent 1-itemsets, denoted  $L_1$ . Then use the 2-itemsets of  $L_1$  to get  $L_2$ , and then  $L_3$ , and so on, until you cannot find the frequent k-itemsets. Apriori algorithm mainly consists of the following three steps:

- (1) Connecting step: connecting k- frequent itemsets to generate (k+1)-candidate sets, denoted by  $C_{k+1}$ . The connect condition of the connecting step is that the two k-itemsets have the same first (k-1) items and different k-th items. Denote  $l_i[j]$  is j-th item of  $l_i$ , the condition is:

$$l_1[1] = l_2[1] \wedge l_1[2] = l_2[2] \wedge \dots \wedge l_1[k-1] = l_2[k-1] \wedge l_1[k] \neq l_2[k]$$

In which  $l_1$  and  $l_2$  are k-item subset of the set collection  $L_k$ ,  $l_1[k] \neq l_2[k]$  is to ensure not to generate duplicate k- itemsets. Itemsets generated by the  $l_1$  and  $l_2$  connection as follows:

$$\{l_1[1], l_1[2], l_1[3], \dots, l_1[k], l_2[k]\}$$

- (2) Pruning step: To pick out the true frequent itemsets  $L_{k+1}$  from the candidate set  $C_{k+1}$ . Because the candidate set  $C_{k+1}$  is the superset of the true frequent itemsets  $L_{k+1}$ . According to the nature of Apriori: any subsets of frequent set must also be frequent, that is any (k-1)- items subsets of k-items must also be frequent. With this property we can find out if the k- items subsets of  $C_{k+1}$  are in  $L_k$ , if not, then remove the candidate (k + 1) - itemset is removed from the  $C_{k+1}$ .
- (3) Counting step: scanning the database, accumulate the number of candidates appearing in the database. If the appear times of a candidate is less than the given minimum support threshold, the candidate itemset will be removed.

## 2.2. Shortage of Apriori algorithm

Apriori is one of the most classical algorithms for mining association rules, but it also has the shortage of low efficiency. The time Apriori algorithm consumes lies mainly in the following three aspects:

- (1) In connection step, when connects  $k$ -itemsets to generate  $(k+1)$ -itemsets, it compares too many times to determine if the itemsets meets the connection conditions. When  $L_k$  has  $m$   $k$ -itemsets, the time complexity of the connection step is  $O(k*m^2)$ .
- (2) In the pruning step, when determine if a subset of candidate set  $C_{k+1}$  is in the frequent set  $L_k$ , the best situation is to simply scan once to get the result, while the worst-case is that it needs to scan  $k$  times to find that the  $k$ -th subset of  $C_{k+1}$  is not in the  $L_k$ . So the average times need to scan and compare the  $L_k$  is  $|C_{k+1}| * |L_k| * k / 2$ .
- (3) In counting step, when accumulate the support times of itemsets in  $C_{k+1}$ , we need to scan the database for  $|C_{k+1}|$  times.

Taking into account these three aspects of time-consuming steps of classical Apriori algorithm, this article presents an improved Apriori algorithm based on prefix-itemset.

## 3. IMPROVED APRIORI ALGORITHM

### 3.1. Improved Apriori algorithm

In 1.2 we have analyzed the shortcomings of classical Apriori algorithm, so its improvements also focus on the three steps mentioned in 1.2. Since the records are already sorted by the dictionary, therefore the candidate set generated by Apriori algorithm is ordered.

- (1) Prefixed-itemset-based storage

In the improved algorithm we proposed a new method to store the itemsets. For each itemset in  $L_k$ , we use a structure similar to Map <key, value> to store them, in which we save the forward  $(k-1)$ - item content as the key while the last item content as the value. After having all the itemsets saved in the new format, we group all the itemsets with the same key and store the union of their values as the new value.

For example: the database is shown in Table 1 and the minimum support is 2.

Table 1 database

TID	Itemset
T1	A,B,E
T2	B,D
T3	B,C
T4	A,B,D
T5	A,C
T6	B,C
T7	A,C
T8	A,B,C
T9	A,B,C,E

The traditional Apriori algorithm will scan the database to obtain the times each item appears in the database, to form the 1- itemsets, and then to generate the 2-itemsets that meets the minimum support, that is 2. Here is the content generated by the classical Apriori algorithm.

Table 2 classical Apriori algorithm

1-itemset		2- itemset	
Item	Count	Item	Count
A	6	AB	4
B	7	AC	4
C	6	AE	2
D	2	BC	4
E	2	BD	2
		BE	2

While Table 3 shows how we store the itemsets with the prefix-itemset-based storage.

Table 3 prefix-itemset-based storage

	Prefixed-key	Value
1-itemset	NULL	{A, B, C, D, E}
2-itemset	A	{B, C, E}
	B	{C, D, E}

As shown in Table 3, 1- itemset has only one item, so the key of 1- itemset is NULL. Besides, we can infer the length of the itemset from the length of the key because the length of the value of the key stores all the items in the itemset but the last item.

### (2) Prefixed-itemset-based connecting step

After the establishing of prefix-itemset-based storage, when we have to generate (k+1)- itemset by connecting the two k- itemsets, we can simply combine two different items in the value, and then generate new itemset with the key. For example, when connecting the 2-itemset with the prefix-key of A in Table 3, we can generate the 3- itemset by combine the value and get the result as  $\{\{B, C\}, \{B, D\}, \{C, D\}\}$ .

### (3) Prefixed-itemset-based pruning step

In chapter 1.1 we know that (k+1)- itemsets are generate from two k- itemsets, and if any k- itemset subset of the (k+1)-itemset does not exist in  $L_k$ , then we have to remove the (k+1)-itemset from  $C_{k+1}$ .

**Theorem:** If we generate a (k+1)-itemset by connecting two k-itemsets,  $l_1$  and  $l_2$ , and one k-itemset of all the k-itemset subset does not exist in  $L_k$ , then the k-itemset subset must contains both  $l_1[k]$  and  $l_2[k]$ .

**Prove:** Assume  $l_1$  and  $l_2$  are both  $k$ -itemset, and the  $(k+1)$ -itemset generated by connecting  $l_1$  and  $l_2$  is  $\{l_1[1], l_1[2], l_1[3], \dots, l_1[k], l_2[k]\}$ . If the  $k$ -itemset does not contain both  $l_1[k]$  and  $l_2[k]$ , then the possible options are  $\{l_1[1], l_1[2], l_1[3], \dots, l_1[k]\}$  and  $\{l_1[1], l_1[2], l_1[3], \dots, l_2[k]\}$ , that is  $l_1$  and  $l_2$ , and both  $l_1$  and  $l_2$  come from  $L_k$ , so if the  $k$ -itemset does not belong to  $L_k$ , then it must contain both  $l_1[k]$  and  $l_2[k]$ .

So in prefixed-itemset-based pruning step, we can simply consider the subset of  $(k+1)$ -itemset which contains both the last two items. With the example from Table 3, we can get the result as follow.

Table 4 pruning step

Subset of 3-itemset	If belong to $L_2$
B,C	yes
B,E	yes
C,E	no
C,D	no
C,E	no
D,E	no

As shown in Table 4, only  $\{\{B,C\}, \{B,E\}\}$  are possible 2-itemset subsets, plus the corresponding prefix-key, that is  $\{\{A,B,C\}, \{A,B,E\}\}$ , namely the candidate set  $C_3$ .

After the pruning step, we have to scan the database to accumulate the times the itemset appears. After accumulating the times of itemsets after pruning step, we can find that both  $\{A,B,C\}$  and  $\{A,B,E\}$  meet the minimum support, and then we add them to the prefix-itemset-based storage as follows.

Table 5 3-itemset storage

	prefix-key	Value
3-itemset	A,B	$\{C,E\}$

### 3.2. Algorithm

The algorithm is described as follow:

Input : Database D, minimum support  $\min\_sup$

Output : frequent itemsets  $L$

1)  $L_1=1$ -itemset of D

2)  $\text{Map}\langle\text{String}[], \text{String}[]\rangle$  map;

3) Import  $L_1$  to map, set the key as null, value as the union of items in  $L_1$

4) for( $k=2; L_{k-1} \neq \phi; k++$ ) {

5)  $C_k = \text{pre\_apriori\_gen}(\text{map}, k-2)$ ;

6) count the appear times of every itemset of  $C_k$ ,  $L_k = \{c \in C_k | c.\text{count} > \min\_sup\}$

7) }

8) Return  $L_k$ ;

procedure  $\text{pre\_apriori\_gen}(\text{map}: \text{Map}\langle\text{String}[], \text{String}[]\rangle; k: \text{int})$

1) for each key in map {

```

2)   if(key.length()==k){
3)       c:=key plus two items from value
4)       if(map.containsKey(c[0:k])){
5)           If( any (k+1)- itemset subset belong to (key,value)){
6)               put c into Ck
7)           }
8)       }
9)   else continue ;
10) }
11) Return Ck

```

#### 4. EXPERIMENT AND RESULTS

The data of the experimental is a total of 120000 patients with diabetes clinical prescription data in Ruijin Hospital, the data records the prescription drug number per user per visit. This experimental machine is configured to Core i5 2.7GHz 8GB processor, 1866MHz LPDDR3 Intel memory.

This experiment compares the classical Apriori algorithm and the prefixed-itemset-based algorithm from two aspects, one is to compare the operation efficiency with fixed total tests and variable minimum support, the other is to compare the operation efficiency with fixed minimum support and variable total tests.

The result of the first experiment is as follows.

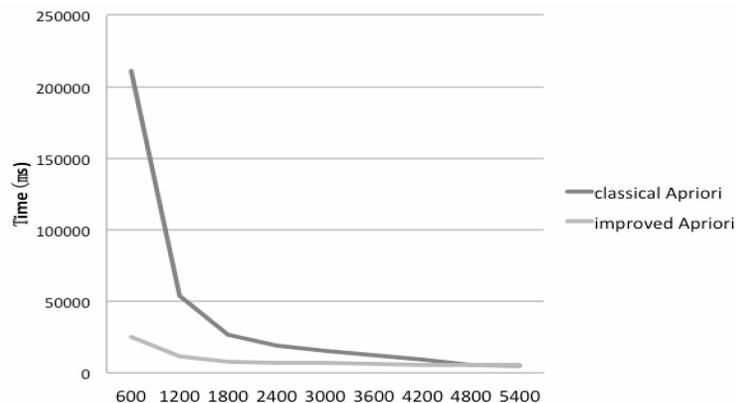


Figure 1. Time consuming with variable min\_sup

The picture above shows the time consuming of the two algorithms when given fixed total test and variable min\_sup. We can infer that the less min\_sup is, the more operation efficiency the improved algorithm improves. And when the min\_sup increases to a certain point, the classical Apriori algorithm and the improved algorithm are of the same efficiency.

Table 6 improvements under variable min\_sup

Min_sup (total 12w)	Classical Apriori Time(ms)	Improved Apriori Time(ms)	Improvement (%)
600	210696	25192	81.44%
1200	53822	11648	68.63%
1800	26317	7614	51.88%
2400	19359	7127	38.99%
3000	15508	6753	56.45%
3600	12393	5842	52.86%
4200	9017	5424	39.85%
4800	5705	5175	9.29%
5400	4868	5161	-6.02%

Table 6 shows the specific operation time of the classical Apriori algorithm and the improved Apriori algorithm and the comparison between them.

The results of the second experiment are as follows.

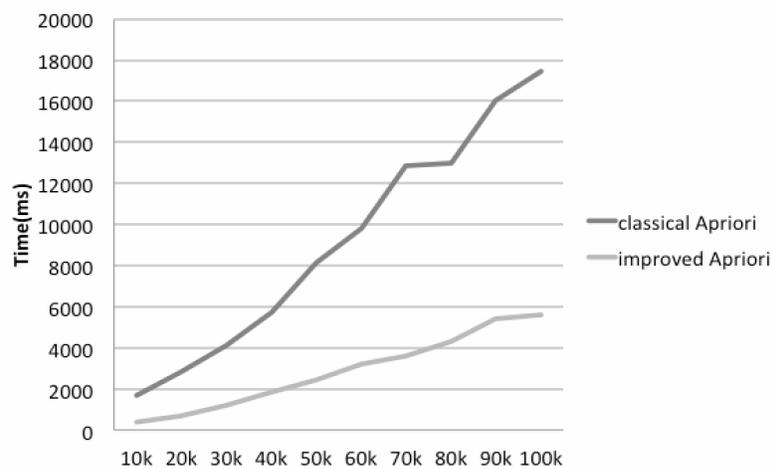


Figure 2. Time consuming with variable total test

The picture above shows the time consuming of the two algorithms when given fixed min\_sup and variable total test. And we can tell that when the total test becomes larger, the improvements become more obvious.

Table 7 improvements under variable total test

Total tests (min_sup% =2%)	Classical Apriori Time(ms)	Improved Apriori Time(ms)	Improvement (%)
10k	1686	390	76.87%
20k	2839	695	75.52%
30k	4088	1229	69.94%
40k	5729	1846	67.78%
50k	8141	2409	70.41%
60k	9833	3197	67.49%
70k	12848	3630	71.75%
80k	13004	4339	66.63%
90k	16007	5442	66.00%
100k	17438	5588	67.96%

Table 7 shows the specific operation time of the two algorithm and we can learn from the table that when the min\_sup is fixed to 2% of the total test, the improvement rate is about 70%.

Experiments have shown that the prefix-itemset-based Apriori algorithm is effective and feasible.

#### 4. SUMMARY

In this paper, we described the Apriori algorithm specifically, and pointed out some limitations of the classical Apriori algorithm during the two steps of the algorithm, namely the connection and the paper cutting steps, and proposed the method of prefixed-itemset-based data storage and the improvements based on it. With the help of prefixed-itemset-based data storage, we managed to finish the connecting step and the pruning step of the Apriori algorithm much faster, besides we can store the candidate itemsets with smaller storage space. Finally, we compare the efficiency of classical Apriori algorithm and improve Apriori algorithm on the aspect of support degree and the total number, and the experimental results on both aspects proved the feasibility of the prefixed-itemset-based algorithm.

**REFERENCES**

- [1] Han J, Kamber M, Pei J, et al. Data mining. Concepts and techniques. 3rd ed[J]. San Francisco, 2001, 29(S1):S103–S109.
- [2] Rakesh Agrawal T, Imielinski, and Arun Swami. Mining association rules between set of items in large databases[J]. In Proceedings Of the Acmsigmod Conference, 1993:207--216.
- [3] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.
- [4] Chun-Sheng Z, Yan L. Extension of local association rules mining algorithm based on apriori algorithm[C]//Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on. IEEE, 2014: 340-343.
- [5] Jia Y, Xia G, Fan H, et al. An Improved Apriori Algorithm Based on Association Analysis[C]//2012 Third International Conference on Networking and Distributed Computing. 2012.
- [6] Shuangyue L, Li P. Analysis of Coal Mine Hidden Danger Correlation Based on Improved A Priori Algorithm[C]//Intelligent Systems Design and Engineering Applications, 2013 Fourth International Conference on. IEEE, 2013: 112-116.
- [7] Wang P, An C, Wang L. An improved algorithm for Mining Association Rule in relational database[C]//Machine Learning and Cybernetics (ICMLC), 2014 International Conference on. IEEE, 2014, 1: 247-252.
- [8] Vaithyanathan V, Rajeswari K, Phalnikar R, et al. Improved apriori algorithm based on selection criterion[C]//Computational Intelligence & Computing Research (ICCIC), 2012 IEEE International Conference on. IEEE, 2012: 1-4.
- [9] Lin X. MR-Apriori: Association Rules algorithm based on MapReduce[C]//Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on. IEEE, 2014: 141-144.
- [10] Zhang, Wenjing, Donglai Ma, and Wei Yao. "Medical Diagnosis Data Mining Based on Improved Apriori Algorithm." Journal of Networks 9.5 (2014): 1339-1345.
- [11] Wu, Huan, et al. "An improved apriori-based algorithm for association rules mining." Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on. Vol. 2. IEEE, 2009.
- [12] Wang, Yuan, and Lan Zheng. "Endocrine Hormones Association Rules Mining Based on Improved Apriori Algorithm." Journal of Convergence Information Technology 7.7 (2012).
- [13] Lin, Ming-Yen, Pei-Yu Lee, and Sue-Chen Hsueh. "Apriori-based frequent itemset mining algorithms on MapReduce." Proceedings of the 6th international conference on ubiquitous information management and communication. ACM, 2012.
- [14] Chai, Sheng, Jia Yang, and Yang Cheng. "The research of improved Apriori algorithm for mining association rules." Service Systems and Service Management, 2007 International Conference on. IEEE, 2007.
- [15] Han J, Pei J, Yin Y. Mining Frequent Patterns without Candidate Generation[J]. Proceeding of Acmsigmod International Conference Management of Data, 1999, 29(2):1-12.

**AUTHORS**

**Yu Shoujian**, the vice professor, main research direction: Web services, enterprise application integration, database and data warehouse;

**Zhou Yiyang**, master, the main research direction: data mining, machine learning

# STATE SPACE GENERATION FRAMEWORK BASED ON BINARY DECISION DIAGRAM FOR DISTRIBUTED EXPLICIT MODEL CHECKING

Nacer Tabib<sup>1</sup>, Jean Michel Ilie<sup>2</sup>, and Djamel Eddine Saidouni<sup>1</sup>

<sup>1</sup>Misc Laboratory, Constantine 2 University , Algeria  
{tabib, saidounid}@misc-umc.org

<sup>2</sup>Lip6 Laboratory, UPMC, France  
{jeanmichel.ilie}@upmc.fr

## **ABSTRACT**

*This paper proposes a new framework based on Binary Decision Diagrams (BDD) for the graph distribution problem in the context of explicit model checking. The BDD are yet used to represent the state space for a symbolic verification model checking. Thus, we took advantage of high compression ratio of BDD to encode not only the state space, but also the place where each state will be put. So, a fitness function that allows a good balance load of states over the nodes of an homogeneous network is used. Furthermore, a detailed explanation of how to calculate the inter-site edges between different nodes based on the adapted data structure is presented.*

## **KEYWORDS**

*Graph distribution, Binary Decision Diagram, State space generation, Formal verification, Model Checking.*

## **1. INTRODUCTION**

An efficient way to improve applications' performances is to use networks. In fact, many already existent applications have been transformed from their simple versions to distributed ones whether they are not initially implemented in a distributed version in the aim of increasing the storage capacity and driving the computing more quicker.

Let's take the formal verification [1] of systems as an example of such applications. An attractive solution to face the major problem of these applications which focus on the combinatorial states space explosion and computing time is the distribution of the graph (states space)[2].

Despite the large use of graphs [3] in computing science domains, they still meet so serious and heavy difficulties especially when certain thresholds and limits are exceeded. That is why it is useful to split the main graph into a set of distributed sub-graphs.

The workload balancing, minimization of the distributed inter-site communication of an unreliable network represent two important factors that are necessary to take them into account in order to generate an ideal distribution of the graph. Both of them influence the application's performances and because of this reason, taking them into account makes the graph distribution a really hard task.

Using several computers of small capacities all together would give an unlimited capacity in term of speed and memory. However, the main inconvenient of distributed algorithms is on distributing the states space of the graph without taking into account the workload balancing that will affect directly the distributed verification application's performances. Besides considering the workload balancing and the distributed inter-nodes edges separately are not enough to improve the distributed verification performances [4].

Several solutions have been proposed to tackle this problem such as equivalence relations, partial order based relations [5] [6]. Although, these solutions reduce the graph size significantly, the memory capacity remains a problem when dealing with very complex systems.

Nowadays, workstations clusters give more and more hardware resources availability, hence we can represent large graph over the cluster where each workstation can hold a sub-graph [7] [8]. But most works use either the symbolic methods based on BDD [9], [10] or explicit methods [7]. A new approach of distributing system states space is proposed in this paper. This new framework developed is based on a compressed format of data structure called Distribution with Binary Decision Diagram (DBDD) to keep a local vision of the whole system. The framework exposes through its API a set of services that can be used by distributed algorithms in order to distribute graphs and perform a distributed verification.

The paper is organized as follows. In Section 2 we introducing fundamental concepts of distributed graphs, BDD and Petri nets, then we move to present our Approach through different subsections in the same part. After deeply presenting the algorithm in Section 3, we make some experiments on the algorithm to show its performances comparing to other algorithms of graph distributing in Section 4 and Section 5. Finally, we achieve the paper by Section 6 to conclude. In the following sections we use interchangeably the terms graph and states space, where we mean by states space a graph generated from a Petri net specification representing its behavioural semantics.

## 2. BASIC CONCEPTS

The graph to be distributed is generated from a petri net specification. We briefly recall the definitions of some basic concepts necessary in the following sections.

### 2.1. Distributed Graph

Let  $W = \{W_k\}_{k=1..N}$  be  $N$  sites, a distributed graph (noted DiG), is a graph with a function of distribution (partial)  $f^k$ .

$$DiG = (G, f^k)_{k=1..N}$$

such that :

- $G = (V, E)$  : an oriented graph.

- $f^k: G \rightarrow G_k$  is an application of  $G$  in  $G_k$ , such that  $G_k = (V_k, E_k)$

**Notation 21**  $\{G_k\}_{1 \leq k \leq N}$  is a set of subsets called fragments  $G_k$ , such that  $\cup V_k = V$  and  $\cup E_k \subseteq E$

**Definition 1.** a fragment  $G_k$  is defined by  $G_k = (V_k, E_k)$  such that :

- $V_k \subseteq V: V_k$  is a fragment of nodes of  $V$  in the site  $W_k$ .
- $E_k = E_k^L \cup E_k^R$  such that  $E_k^L \cap E_k^R = \emptyset$  : the set of intra-site and inter-sites edges with :
  - $E_k^L \subseteq V_k^2$  is the set of edges between nodes belonged in the same site  $W_k$  (Local edges).
  - $E_k^R \subseteq V_k \times (V \setminus V_k) = \{(v_k, v'_k) \text{ such that } v_k \in V_k \text{ and } v'_k \notin V_k\}$  : is the set of edges whose the origins are in the local sites and the goals are in the remote sites (Remote edges).
  - $\alpha_k$  and  $\beta_k$  are two applications of  $E_k$  in  $V$  such that for all edges  $e = (v, v') \in E$  :
    - $\alpha_k(e) = v \in V_k$  : indicate the origin of the edge  $e$ .
    - $\beta_k(e) = v' \in V_k$  if  $e \in E_k^L$  and  $\beta_k(e) = v' \notin V_k$  else.

**Notation 2.2** given a set  $S$ ,  $|S|$  denotes its cardinality (the number of elements).

Figure 1 represents a distributed graph over sites (nodes) of a cluster of workstations (workers). We assume that the initial graph is so large that it can't be hold in one machine so distributing it over a different sites while generating it make it possible to take advantage of distributed memory hence we can represent more and more large graphs that correspond to very complex systems.

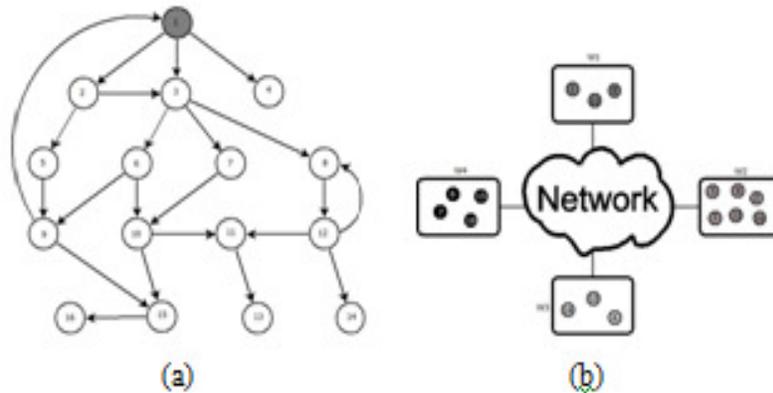


Fig.1. graph before distribution (a) and after (b)

## 2.2. Petri Net Related Definitions

- A Petri net [11] is a tuple  $(S, T, W)$  where  $S$  is the set of places,  $T$  is the set of transitions such that  $S \cap T = \emptyset$ , and  $W : ((S \times T) \cup (T \times S)) \rightarrow N = \{0, 1, 2, \dots\}$  is the weight function. Graphically, transitions of  $T$  are represented by rectangles, places of  $S$  by circles and weight function by arrows associated with their weights. We suppose that all nets are finite, i.e.  $|S \cup T| \in N$ .

- For  $x \in S \cup T$ , the pre-set  $\cdot x$  is defined by  $\cdot x = \{y \in S \cup T \mid W(y,x) \neq 0\}$  and the post-set  $x \cdot$  is defined by  $x \cdot = \{y \in S \cup T \mid W(x,y) \neq 0\}$ .
- The *marking* of a Petri net  $(S,T,W)$  is defined as a function  $M : S \rightarrow \mathbb{N}$ . A marking is generally represented graphically by putting tokens in places.
- Safety-Petri net is a Petri net  $(S,T,W)$  such that for any  $s$  of  $S : M(s) \leq 1$
- The transition rule stipulates that a transition  $t$  is enabled by  $M$  iff  $M(s) \geq W(s,t)$  for all  $s \in S$ . The firing of a transition  $t$  will produce a new marking  $M'$  defined by  $M'(s) = M(s) - W(s,t) + W(t,s)$  for all  $s \in S$ . The occurrence of  $t$  is denoted by  $M \xrightarrow{t} M'$ .
- Two transitions  $t_1$  and  $t_2$  (not necessarily distinct) are concurrently enabled by a marking  $M$  iff  $M(s) \geq W(s,t_1) + W(s,t_2)$  for all  $s \in S$ .
- A marked Petri net  $(S,T,W,M)$  is a Petri net  $(S,T,W)$  with an initial marking  $M$ .
- An alphabet  $A$  is a finite set; we suppose that  $\tau \in A$  ( $\tau$  will indicate invisible action, or *silent action*).
- The labeling of a Petri net  $N = (S,T,W)$  is a function  $\lambda : T \rightarrow A \cup \{\tau\}$ . If  $\lambda(t) \in A$  then  $t$  is said to be *observable* or *external*; at the opposite,  $t$  is *silent* or *internal*.
- $\Sigma = (S,T,W,M,\lambda)$  is a labeled system iff  $(S,T,W,M)$  is a marked Petri net and  $\lambda$  is a labeling function of  $(S,T,W)$ .

### 2.3 BDD

A Binary Decision Diagram or BDD [10] is data structure used for representation of Boolean functions in the form of rooted directed acyclic graph. A BDD is a rooted directed acyclic graph  $G = (V,E)$  with node set  $V$  containing two kinds of nodes, *non-terminal* and *terminal* nodes (Figure 2). A non-terminal node  $v$  has as tag a variable  $index(v) \in \{x_1, x_2, \dots, x_n\}$  and two children  $low(v)$ ,  $high(v) \in V$ . The final nodes are called *0-final* and *1-final*. A BDD can be used to compute a Boolean function  $f(x_1, x_2, \dots, x_n)$  in the following way. Each input  $a = (a_1, a_2, \dots, a_n) \in \{0, 1\}^n$  defines a computation path through the BDD that starts at the root. If the path reaches a non-terminal node  $v$  that is labelled by  $x_i$ , it follows the path  $low(v)$  if  $a_i = 0$ , and it follows the path  $high(v)$  if  $a_i = 1$ . The label of the terminal node determines the return value of the BDD on input  $a$ . the BDD is called "ordered" if the different variables appear in the same order on all the ways from the root (Figure 2).

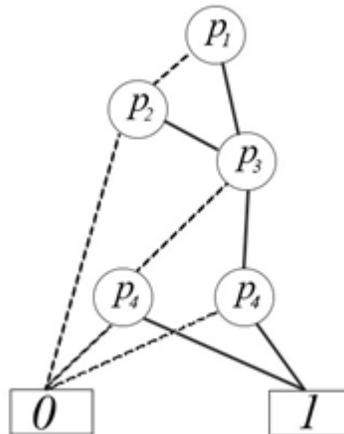


Fig.2. Binary decision diagram

**Generating a BDD from a Petri Net BDD:** can represent a state space generated from a safe petri Net in an efficient high compressed format. The Figure 3(b) represents a BDD generated from a safe Petri Net 3(a). It uses a set of variables proportional to the number of places in petri net in this example it uses 6 variables to code the different configurations of petri net  $p1, p2, p3, q1, q2$  and  $q3$ .

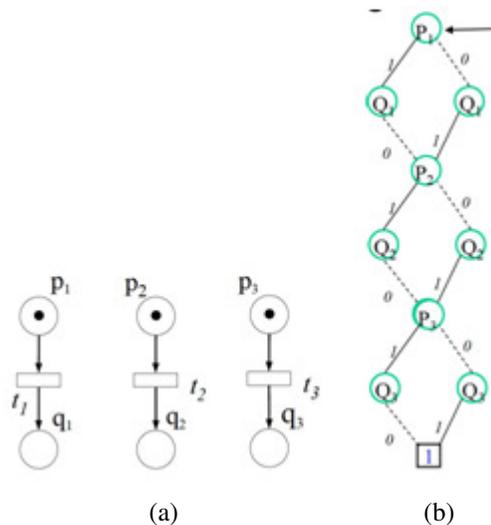


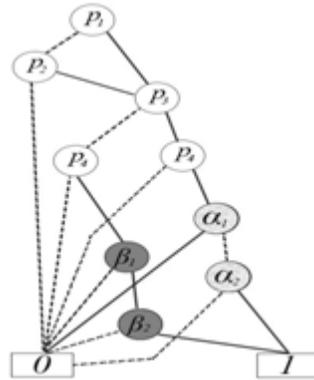
Fig.3. Petri net specification (a) and corresponding BDD (b)

### 3. PROPOSED APPROACH

Here we are going to present a new framework for graph distribution based on adapted data structure called (DBDD) Distribution with Binary Decision Diagram, the framework provide functions that can be used by parallel and distributed algorithms to generate an explicit state space or to get the location of specific states successors in the distributed graph. Hence the DBDD represent a global state of the system which decrease the communication between several nodes of the network workers and ensure a better fault tolerance.

### 3.1. Sites Encoding

The DBDD in addition to representing the reachability graph of petri net it encodes the place of each state by injection of a additional game of variables, each variable represent the site where the state is meant to be. Figure 4 represents an example of the encoding of two sites by adding variables which represents these two site ( $\alpha_1, \alpha_2$ ) to encode the first site in binary (01). and ( $\beta_1, \beta_2$ ) for the second site (10).



ig.4. DBDD represents a graph distributed over two nodes

### 3.2. DBDD generation

Algorithm 1 below represents the generation of the DBDD, variables are chosen according a binary variable  $bddSite$ . The fitness function  $F$  ensures a good load balance.

---

**Algorithm 1:** *GenerateDBDD(PNet)*

---

**Constants:**  $bdd0$  : bdd representing the site0 (00);  
 $bdd1$  : bdd representing site1 (01);  
 $bdd2$  : bdd representing site2 (10)  
 $bdd3$  : bdd representing site3 (11);  
 $s_0$  : bdd representing the first marking  $M_0$  of

PNet

$T$  : bdd representing the transition relation

**Variables:**  $Todo = \{s_0\}$  : set of states to be processed ;  
 $Visited = \{\}$  : set of processed states;

```

1 while  $Todo \neq Visited$  do
2    $bddSite \leftarrow \max(F)(\{bdd0, bdd1, bdd2, bdd3\})$ ;
3    $Visited \leftarrow Todo$ ;
4    $Todo \leftarrow Todo \cup T(Visited) \cup bddSite$ 
5 end

```

---

### 3.3. Fitness function

The site to be chosen for a given set of states is calculated based on the following fitness function:

$$F = \prod_{i=1}^{i=n} |V_i|$$

In an homogeneous network all Sites have the same memory capacity, and a good balance load is when each site hold exactly  $\frac{|V|}{n}$  such that  $\sum_{i=1}^{i=n} |V_i| = |V|$

#### 4. IMPLEMENTATION AND EVALUATION

The proposed approaches are implemented with JavaBDD [12] (An open source library for manipulating BDD, it is also a wrapper for other libraries such Buddy [13] and Cudd [14]) tested on a network of PC with a 3.0 GHZ processor and 512 MB of memory. We developed a tool that generates distributed graphs associated to petrinets specifications (Figure 5) which is part of FOCOVE framework.

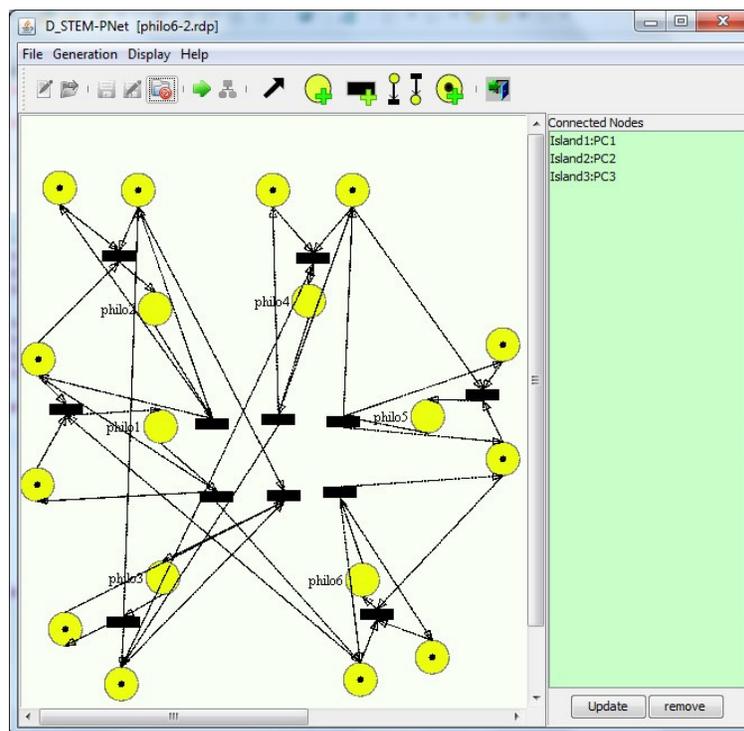


Fig.5. Tool for editing petrinets and generation of state space

#### 5. RESULT AND EXPERIMENTATION

To see the contribution and the advantage of the proposed approach, we compare it to hash function (MD5)[8] based algorithm. Taking examples studied in literature enables us to get more closely to the problem of combinatorial explosion. In the context, we have selected three well known classic case studies in system models. These models include dining philosophers system [15], Peterson solution for mutual exclusion [16] and shared memory system [17].

Table 1. Comparative results of the bdd approach,MD5 based algorithm.

<b>5 sites</b>	<b> V </b>	<b> E </b>	$\sigma_v$ <b>MD5</b>	$\sigma_v(\%)$ <b>MD5</b>	$\sigma_v$ <b>BDD</b>	$\sigma_v(\%)$ <b>BDD</b>
philosophiers	729	3402	21.46	2.9	14.36	1.97
Shared memory	8019	52974	249.61	3.11	96.01	1.19
Peterson	20754	62262	588.67	2.83	607	2.9

The table(1) shows the statistic results according to philosophers, shared memory and Peterson models knowing that the states space has been distributed over 5 sites. The standard deviation of the number of states on each site noted by  $\sigma_v(\%)$  is calculated as follows  $\sigma_v(\%) = \frac{\sigma_v}{|V|}$ . The smaller is the standard deviation  $\sigma_v$ , the better is the distribution over sites, because a tiny  $\sigma_v$  means that the states space is well distributed on the different sites and we see that on table(1). Using the new proposed approach makes it possible to have a fewer  $\sigma_v$  than the one obtained by using the (MD5) based algorithm except for Peterson and this is due to the replication of some states over the sites.

## 6. CONCLUSION

In this paper, we have presented a new framework based on binary decision diagrams algorithm to solve the graph distribution problem in context of formal verification. We have used an adapted data structure which ensures a high compression property, the balance load and fault tolerance. We have also compared our work with md5 based algorithm. Results are promising.

To put in practice the result of this work, an optimization algorithm such as evolutionary algorithm or local search may be applied to improve the inter-site communication and tackle also with the variable order problem in BDD. Beside this, different verification algorithms may be applied on the distributed graph generated to verify properties of complex systems.

## REFERENCES

- [1] Edmund M Clarke, Orna Grumberg, & Doron Peled. Model checking. MIT press,(1999).
- [2] Antti Valmari(1998). The state explosion problem , Lectures on Petri nets I: Basic models, pp 429–528. Springer.
- [3] Douglas Brent West et al (2001). Introduction to graph theory, volume 2. Prentice hall Upper Saddle River.
- [4] Hans Hansson & Bengt Jonsson(1990). A calculus for communicating systems withtime and probabilities, In Real-Time Systems Symposium, 1990. Proceedings., 11th, pp 278–287.
- [5] François Vernadat, Pierre Azéma, & François Michel(1996). Covering step graph , Application and theory of Petri nets, pp 516–535. Springer.
- [6] Patrice Godefroid, J van Leeuwen, J Hartmanis, G Goos, & PierreWolper. Partialorder(1996) methods for the verification of concurrent systems: an approach to the stateexplosion problem.
- [7] Hubert Gavel, Radu Mateescu, & Irina Smarandache(2001). Parallel state space construction for model-checking. , Model Checking Software, pp 217–234. Springer.

- [8] Hubert Garavel, Radu Mateescu, Wendelin Serwe(2013), et al. Génération et manipulation d'espaces d'états distribués avec cadp: expériences sur grid'5000, Conférence en Parallélisme, Architecture et Système ComPAS'2013.
- [9] Stefan Blom & Simona Orzan(2003). Distributed branching bisimulation reduction of state spaces. Electronic Notes in Theoretical Computer Science, vol.1 n- 89 pp 99–113.
- [10] Randal E Bryant.(1992 ) Symbolic boolean manipulation with ordered binary-decision diagrams. ACM Computing Surveys (CSUR), vol.3 n° 24 pp 293–318.
- [11] Eike Best & Harro Wimmel (2013 ). Structure theory of petri nets, Transactions on Petri Nets and Other Models of Concurrency VII, pp 162–224. Springer.
- [12] <http://javabdd.sourceforge.net/>
- [13] <http://sourceforge.net/projects/buddy/>
- [14] <http://vlsi.colorado.edu/~fabio/CUDD/cuddIntro.html>.
- [15] "NetLogo Models Library: Sample Models/Computer Science Standards"  
<http://ccl.northwestern.edu/netlogo/models/DiningPhilosophers>
- [16] "Model Checking Contest, "Peterson model" [http://sumo.lip6.fr/Peterson\\_model.html](http://sumo.lip6.fr/Peterson_model.html)
- [17] "Model Checking Contest, "Shared memory model" [http://sumo.lip6.fr/SharedMemory\\_model.html](http://sumo.lip6.fr/SharedMemory_model.html)

*INTENTIONAL BLANK*

# A NEW ALGORITHM FOR CONSTRUCTION OF A P2P MULTICAST HYBRID OVERLAY TREE BASED ON TOPOLOGICAL DISTANCES

Sergej Alekseev<sup>1</sup> and Jörg Schäfer<sup>2</sup>

<sup>1</sup>Department of Computer Science and Engineering, Computer Networks and OS,  
Frankfurt University of Applied Sciences, Germany  
alekseevf@fb2.fra-uas.de

<sup>2</sup>Department of Computer Science and Engineering, Distributed Systems,  
Frankfurt University of Applied Sciences, Germany  
jschaefer@fb2.fra-uas.de

## **ABSTRACT**

*In the last decade Peer to Peer technology has been thoroughly explored, because it overcomes many limitations compared to the traditional client server paradigm. Despite its advantages over a traditional approach, the ubiquitous availability of high speed, high bandwidth and low latency networks has supported the traditional client-server paradigm. Recently, however, the surge of streaming services has spawned renewed interest in Peer to Peer technologies. In addition, services like geolocation databases and browser technologies like Web-RTC make a hybrid approach attractive.*

*In this paper we present algorithms for the construction and the maintenance of a hybrid P2P overlay multicast tree based on topological distances. The essential idea of these algorithms is to build a multicast tree by choosing neighbours close to each other. The topological distances can be easily obtained by the browser using the geolocation API. Thus the implementation of algorithms can be done web-based in a distributed manner.*

*We present proofs of our algorithms as well as practical results and evaluations.*

## **KEYWORDS**

*Distributed algorithms, peer-to-peer (P2P), hybrid, overlay multicast tree, live streaming*

## **1. INTRODUCTION**

Peer-to-Peer (P2P) streaming has become more and more popular nowadays again after interest in general P2P has generally decreased after the initial enthusiasm in the late 90 – partially due to the ubiquitous and quick availability of high speed, high bandwidth and low latency networks which has supported the traditional client-server paradigm in the last decade. The central strength of P2P streaming systems is the capability of sharing resources so that larger (and more costly) servers can be replaced by smaller (and cheaper) computers. The P2P networks are build usually as a logical overlay network. The contribution of this paper is the construction and management of a P2P multicast tree streaming overlay where the nodes are physically close to each other in

the underlying network. In this paper we present two algorithms. The first is the joining algorithm that each node runs when it enters the system. The essential idea of the algorithm is to construct a multicast tree structure by finding a suitable neighbour in the overlay multicast tree and considering resources of peers. The second algorithm handles a host leaving that occurs gracefully or accidentally. For both algorithms we provide full mathematical proofs of minimality features. In addition, we present some experimental results and evaluations. And finally we conclude our paper with remarks on possible future work.

## 2. RELATED WORKS

In recent years a number of P2P-based applications for stream delivery have been developed – e.g. Zattoo (<http://zattoo.com>), PPTP (<http://www.pptv.com>) and Octoshape (<https://octoshape.com>).

To improve the scalability and to optimise the usage of resources in the P2P network, several approaches have been proposed. In [1] various problems that arise due to the fact of P2P systems being highly dynamic and heterogenous are examined. It focuses especially on resilience mechanisms. In [2] and [6] an overview of application and network layer mechanisms are presented and the Mesh and Multiple-Tree P2P overlays are compared.

Several applications have been developed for various categories of mesh based P2P streaming. The authors of [8] and [7] present a hybrid approach for overlay construction and data delivery in an application-layer multicast. The HyPO approach in [7] optimizes the overlay by organizing peers with similar bandwidth ranges in a geographical area into a mesh overlay. The ToMo approach in [7] combines the strong points of a tree-based structure and a mesh-based data delivery to a two-layer hybrid overlay. The mTreebone of [9] is a collaborative tree-mesh design that leverages both mesh and tree structures. The key idea is to identify a set of stable nodes to construct a tree-based backbone with most of the data being pushed over this backbone. AnySee [5] is a mesh based P2P system in which resources are assigned based on their locality and delay.

In the present work we propose algorithms to construct a tree based multicast overlay based on topological distances. Similar approaches are described in [12], [3] and [14]. Already in [20] an architecture has been proposed for designing a global internet host distance estimation service. However, only relatively recently geographical information has become practically available from freely available geolocation databases [16], and therefore ideas which have been of theoretical value only have now become practical, see also [19]. The approach used in [12] and [3] organizes the peers into a hierarchy of clusters such that the neighboring peers are grouped into the same cluster. The overlay network is build from the cluster leaders to the other members recursively. In [14] a locality-aware P2P overlay construction method, called Nearcast, is proposed which builds an efficient overlay multicast tree by letting each peer node choose physically closer nodes as its logical children. Whereas there is rather comprehensive coverage of theoretical P2P algorithms and mathematical theorems on some of them like e.g. the T-Man protocol, see [4], up to our best knowledge, no minimality results have been proven for the overlay networks like the one described above but rather simulation results have been computed. In our work we propose algorithms which minimise the routing costs, usage of peer resources and end-to-end delay based on the topological location of peers. We provide a proof for the minimality of routing costs and provide evidence for keeping end-to-end delay low.

### 3. PROPOSED APPROACH

The concept of P2P multicasting [11], [12] is often applied to reduce the costs needed to deploy and to maintain services related to streaming of various content to many users, e.g. VoD, IPTV, radio, news channels, etc. In this paper, we propose an approach to the construction of a P2P overlay multicast tree with the goal to solve the following important problems:

- **Optimal routing between peers:** Transmission at an overlay P2P-network might be inefficient, especially when the P2P-network is randomly constructed. This stems from the fact that the distance between peers physically or topologically is not considered by constructing the P2P-network.
- **Optimal usage of peer resources:** Peer resources include available bandwidth, processing power and storage space.
- **End-to-End delay:** The end-to-end delay is the latency, accumulated peer by peer, for the delivery of a data packet along the overlay path from the source host to an end host. To reduce this delay the height of the multicast tree should be kept small.
- **Handling of peer connections:** In practice the P2P-network need to deal with peers joining the network and peers that leave voluntarily or due to failure.

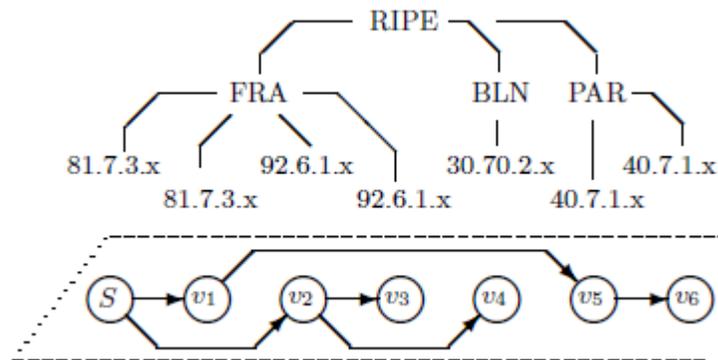


Figure 1. Topological search tree and p2p multicast tree structure.

To overcome these problems, we propose algorithms for the construction and the management of an overlay P2P-network. Our algorithms use the topological distances between peers to guarantee the optimal routing costs. We define two data structures, a topological search tree and a P2P multicast tree (fig. 1). The search tree is used to find the nearest peer to be attached to the multicast overlay. This is a special case of the Nearest Neighbour Search (NNS) or closest point search problem. Donald Knuth named this problem the post office problem [10]. The problem relates to an application of the assignment to the next post office. In our case the problem is reduced to the search in the tree and adapted for the search of an optimal usage of peer resources. The P2P multicast tree is used for the actual data transfer.

## 4. P2P OVERLAY MULTICAST TREE CONSTRUCTION AND MAINTENANCE

### 4.1. Definitions and Preliminaries

To identify the topological position of hosts in a network, a unique  $H$ -dimensional coordinate  $C$  is assigned to each host. The idea to use the network coordinates is based on considerations from [13], [14] and [15]. In contrast to the algorithms presented therein, we use in our approach two data structures: the search tree  $T_s$  for searching the nearest neighbour according the topological position in the network and the multicast tree  $T$  to connect hosts to a P2P overlay multicast network.

The multicast overlay tree is defined as  $T = (V, E)$ , where  $V$  is a set of vertices, which represent the end hosts, and  $E$  is a set of directed edges, which represent data delivery streams between the end hosts.

The search tree  $T_s$  is considered as an  $H$ -layered topological tree. According to the topology of the search tree  $T_s$  for each vertex  $v \in V$  the *network coordinate*  $C(v)$  is defined as follows:

$$C(v) = \{C_{H-1}(v), \dots, C_0(v)\} \quad (1)$$

Similarly to the Nearcast method proposed in [14] we use static *network coordinates* and assign to the vertices (end hosts in the physical network) special geographical meanings. The coordinates  $C_{H-1}(v), \dots, C_0(v)$  represent *Regional Internet Registry, Country, City* and  $n$ -th bits of an *IP*-address respectively e.g:

$$\begin{aligned} &\{RIPE, DE, HH, 80.x.x.x, 80.6.x.x., 80.6.60.x\} \\ &\{ARIN, US, NIC, 10.x.x.x, 10.7.x.x., 10.7.50.x\} \end{aligned}$$

The geographical information can be easily obtained from the freely available geolocation databases [16] by using the programming interfaces described in [17] and [18].

Formally the search tree  $T_s$  can be defined using tuple notation as  $T_s = (V_s, E_s)$ , where

$$V_s = \bigcup C(v) \quad (2)$$

and

$$E_s = \bigcup_{0 < i < (H-1)} \{(C_i(v), C_{i+1}(v))\}. \quad (3)$$

Finally we introduce a hierarchical common network distance  $D$  and last common coordinate  $LCC$ , used by our algorithms. The hierarchical common network distance  $D$  between two vertices  $v_x$  and  $v_y$  with the static network *coordinates*:

$$\begin{aligned} C(v_x) &= \{C_0(v_x), \dots, C_i(v_x), \dots, C_{H-1}(v_x)\} \\ C(v_y) &= \{C_0(v_y), \dots, C_i(v_y), \dots, C_{H-1}(v_y)\} \end{aligned}$$

is the number of coordinates with different values and is denoted as  $D(v_x, v_y)$ . Formally the hierarchical common network distance is defined as:

$$\begin{aligned}
D(v_x, v_y) &= H - 1 - m \\
m &= \max_{0 \leq i \leq H-1} \{i \mid C_k(v_x) = C_k(v_y) \forall k \leq i\}
\end{aligned} \tag{4}$$

e.g. for the following vertices

$$\begin{aligned}
v_x &= \{RIPE, DE, FRA, 80.x.x.x, 80.70.x.x\} \\
v_y &= \{RIPE, DE, BLN, 90.x.x.x, 90.80.x.x\}
\end{aligned}$$

the hierarchical common network distance is  $D = 3$ .

The last common coordinate *LLC* of two vertices is the last identical coordinate in the order of  $C_0, C_1, \dots, C_i$ , formally:

$$LCC(v_x, v_y) = C_i \iff C_k(v_x) = C_k(v_y) : 0 < k < i \tag{5}$$

In the example above the  $LCC(v_x, v_y) = DE$ .

## 4.2. Joining algorithm

To construct a multicast overlay tree the joining algorithm connects the hosts to an overlay network by analysing the geolocation information provided by the end hosts. The algorithm can be implemented in a centralized or a distributed manner. The pseudocode of the joining algorithm is shown in fig. 2.

---

**Algorithm JOIN** (*new*,  $T_s$ )

---

- 1  $T_s = T_s \cup C(\text{new})$
- 2 find the nearest neighbour  $n$  of *new* in  $T_s$ ;
- 3 attach *new* to  $n$ ;

---

Figure 2. The pseudocode of the JOIN algorithm

Figure 3 illustrates an example of the joining nodes to an existing overlay network. Initially the overlay multicast tree contains only a source host  $S$  and the search tree  $T_s$  includes the coordinates  $C(S)$  (fig. 3a). To attach the new node  $v_1$  the joining algorithm extends the search tree  $T_s$  by the adding the coordinates  $C(v_1)$  and determines the nearest host by traversing the search tree  $T_s$ . The nearest neighbour can be easily found by a simple tree traversing in  $O(\log n)$  time. The new host  $v_1$  is attached to the host  $S$  (fig. 3b). The fig. 3c illustrates the attaching of the host  $v_2$  to the multicast overlay tree.

To show that the routing in the constructed multicast tree  $T$  is optimally organised, we assign to each edge  $e = \{v_0, v_1\}$  a topological distance value  $D(e) := D(v_0, v_1)$ , that represents the routing costs between the vertices  $v_0$  and  $v_1$ . It is easy to check that  $D$  satisfies all axioms of a metric which is important for the minimality results presented in the sequel (only the triangle inequality is non-trivial). The sum of all distances  $S(T) = \sum_{e \in E} D(e)$  is the total routing costs in the tree. The lower the value  $S(T)$  is, the less

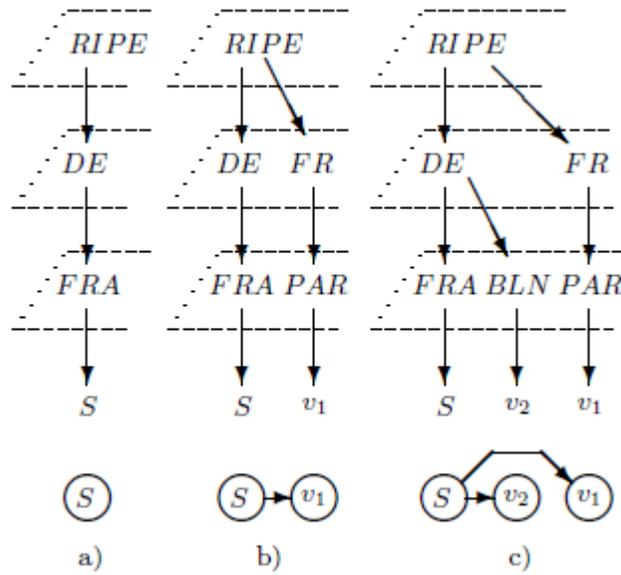


Figure 3. An example of the algorithm *JOIN* execution.

routing overhead is necessary to deliver the content to each vertex in the multicast tree. The topological distance can be thought of a proxy for the “real” distance measured as End-to-End-Delay or other QoS parameters. In the literature (see e.g. [21]) it has been argued, that the topological distance is a reasonable proxy in practice. As next we show that our algorithm constructs a multicast tree  $T$  with minimal routing costs. In other words it is not possible to construct another multicast tree  $T_1$  with  $S(T_1) < S(T)$ .

**Theorem 1.** *The algorithm JOIN (fig. 2) constructs a tree  $T$  with the minimal  $S(T)$  value.*

*Proof.* The correctness of the algorithm is proved by induction on the number of vertices in  $T$ .

*Base case:*  $T = \emptyset$  or  $|T| = 1$  are trivially minimal.

*Induction step:* Assume that  $S(T)$  is minimal for  $n$  connected vertices. Let  $v_{n+1}$  be the next vertex added to the multicast tree  $T$  and  $v \in T$  is the nearest neighbour of  $v_{n+1}$  (fig. 4a). Let us show that  $S(T) + D(v, v_{n+1})$  is minimal.

Consider any multicast trees  $T'$ , where  $v_{n+1}$  not connected to  $v$ . We will see that there is no tree with  $S(T') < S(T)$ .

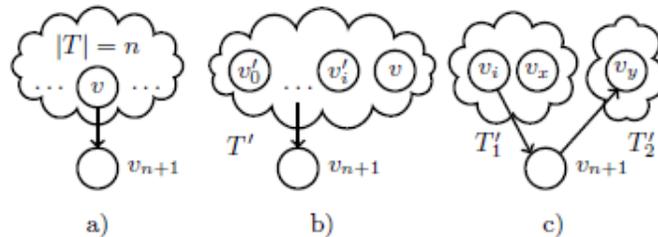


Figure 4. Proof by induction

First assume the vertex  $v_{n+1}$  is connected to any vertices  $v' \in T \setminus v$  as a leaf (fig. 4b). It is easy to see that  $S(T') \geq S(T) + D(v, v_{n+1})$ , because  $D(v', v_{n+1}) \geq D(v, v_{n+1}) \forall v'$  as  $v$  is a nearest neighbour.

Thus, the  $S(T')$  value could only be possibly reduced by connecting the vertex  $v_{n+1}$  such that an edge  $e = \{v_x, v_y\}$  with  $D(v_x, v_y) > D(v, v_{n+1})$  is removed from  $T$  (fig. 4c). Removing an edge from the tree breaks it into two separate subtrees  $T'_1$  and  $T'_2$ . The vertex  $v_y$  is the root of the subtree  $T'_2$ , because it is the successor of the vertex  $v_x$ . In order to connect two subtrees the vertex  $v_y$  must be connected to the vertex  $v_{n+1}$  as a successor and the vertex  $v_{n+1}$  is connected to a vertex  $v_i$  in  $T'_1$ . It is possible that  $v_i = v_x$  or  $v_i = v_n$ , if  $v_n \in T'_1$ .

With  $T' = (T \setminus \{v_x, v_y\}) \cup \{v_{n+1}, v_y\} \cup \{v_i, v_{n+1}\}$ , let us assume the following inequality being strict:

$$S(T') < S(T) + D(v, v_{n+1}) \tag{6}$$

The  $S(T')$  value of  $T'$  can be calculated from the definition as:

$$S(T') = S(T) - D(v_x, v_y) + D(v_{n+1}, v_y) + D(v_i, v_{n+1})$$

As  $v$  has minimal distance, by replacing the  $S(T')$  in the inequality (6) we get:

$$-D(v_x, v_y) + D(v_{n+1}, v_y) + D(v_i, v_{n+1}) < D(v, v_{n+1})$$

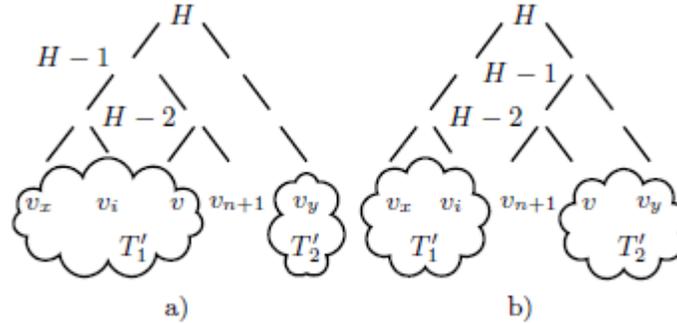


Figure 5. Topological distances

Thus,  $D(v_x, v_y) > D(v_{n+1}, v_y)$  and  $D(v_x, v_y) > D(v_i, v_{n+1})$ .

However, from the definition of topological distances in the search tree  $T_s$  (fig. 5) the following must be true:

$$\begin{aligned} (D(v_x, v_y) = D(v_{n+1}, v_y)) \wedge (D(v_x, v_y) > D(v_i, v_{n+1})) \\ (D(v_x, v_y) > D(v_{n+1}, v_y)) \wedge (D(v_x, v_y) = D(v_i, v_{n+1})) \end{aligned} \tag{7}$$

But (7) contradicts inequality (6). Thus  $S(T) + D(v, v_{n+1})$  is minimal.  $\square$

The algorithm *JOIN* (fig. 2) may construct different trees depending on selection of the nearest neighbour and the order the nodes joining, however the next theorem shows that  $S(T)$  is not affected.

**Theorem 2.** *All trees constructed by the algorithm JOIN (fig. 2) have the same  $S(T)$  value.*

*Proof.* Assume that algorithm JOIN (fig. 2) constructs two different trees  $T_0$  and  $T_1$  for the same set of vertices (end hosts) with  $S(T_0) \neq S(T_1)$ .

According the theorem 1 the value of  $S(T_0)$  and the value of  $S(T_1)$  are minimal. Since  $S(T_0) \neq S(T_1)$ , follows that either  $S(T_0)$  or  $S(T_1)$  is not minimal. So the assumption must be incorrect.  $\square$

The algorithm JOIN (fig. 2) solves the routing costs problem, mentioned in the section 3. However the algorithm does not consider the usage of peer resources. As next we present an extension of the algorithm JOIN to solve the peer capacity problem.

### 4.3 Management of peer-resources

To manage the usage of peer resources the attribute *resource capacity* is assigned to each host in the network model. The *resource capacity* of an end host, denoted by  $R(v)$ , is a maximum number of outgoing links  $e \in E$ , which can be served by the vertex  $v$ . The value  $R(v)$  is calculated based on available bandwidth and other resources of the peer.

The pseudocode of the joining algorithm with the peer-resource management  $JOIN_R$  is shown in fig. 6 (we call  $v$  in  $LCC(new, n)$  iff  $LCC(new, v) = LCC(new, n)$ ). To join a new node the algorithm  $JOIN_R$  finds the nearest neighbour  $n$ , similar to the algorithm JOIN (fig. 2). Instead to attaching the node directly to the nearest neighbour  $n$  found, the algorithms checks all existing hosts with the same topological distance as the vertex  $n$ , whether one of the vertices has enough resources to forward the data link to the new

---

Algorithm  $JOIN_R$  ( $new, T_s$ )

---

```

1   $T_s = T_s \cup C(new)$ 
2  /*  $n$  is a potential neighbour of  $new$  */
3  for (all reachable hosts  $v$  in  $LCC(new, n)$ ) {
4      if( $R(v) \neq 0$ ) {
5          attach  $new$  to  $v$ ;
6           $R(v) = R(v) - 1$ ;
7          return;
8      }
9  }
10 for ( all reachable hosts  $v$  in  $LCC(new, n)$  ) {
11     for (all hosts  $v_x$  connected to  $v$ ) {
12         if( $D(v, new) \leq D(v, v_x)$ ) {
13             insert  $new$  between  $v$  and  $v_x$ ;
14              $R(new) = R(new) - 1$ ;
15             return;
16         }
17     }
18 }
```

---

Figure 6. The pseudocode of the  $JOIN_R$  algorithm

node. For that the last common coordinate  $LCC$  according the definition 5 (section 4.1) is calculated. If one appropriate host  $v$  is found the new node is attached and the resource capacity attribute of the host  $v$  is updated. Otherwise the algorithms checks again all reachable hosts and verifies if the new node can be inserted between a host  $v$  and any hosts connected to  $v$  with  $D(new, v) \leq D(v_x, v)$ .

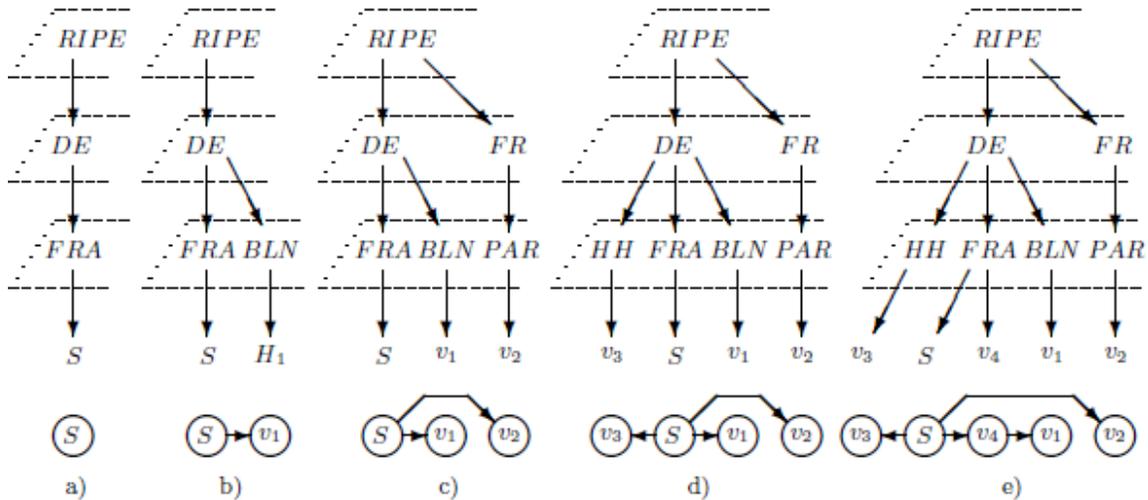


Figure 7. An example of the algorithm  $JOIN_R$  execution

Figure 7 illustrates an example of the algorithm  $JOIN_R$  execution. We define for this example the following condition  $\forall v \in V : R(v) = 3$ . In other words each host is able to maintain three outgoing links.

Initially the overlay multicast tree contains only a source host  $S$  and the search tree  $T_s$  includes the coordinates  $C(S)$  (fig. 7a).

To attach the new node  $v_1$  the algorithm  $JOIN_R$  adds the coordinates  $C(v_1)$  to the search tree  $T_s$  and determines the nearest host of  $v_1$  (fig. 7b). The host  $v_1$  is attached to the host  $S$  and the  $R(S)$  value is updated accordingly  $R(S) = 3 - 1 = 2$ .

Figure 7c illustrates the attaching of the host  $v_2$ . After updating the search tree  $T_s$  the algorithm  $JOIN_R$  checks all potential nearest neighbours, reachable from the  $LLC = RIPE$ . In order to find the  $LLC$ -value, it is enough to traverse backwards the search tree  $T_s$  from the vertex  $v_2$  until the first branch. The potential nearest neighbours of  $v_2$  are the hosts  $S$  and  $v_1$ , because  $D(S, v_2) = D(v_1, v_2) = 2$ . The host  $v_2$  is attached to the host  $S$  and the  $R(S)$  value is updated accordingly  $R(S) = 2 - 1 = 1$ .

The attaching of the host  $v_3$  (fig. 7d) is similar to the previous step. The  $R(S)$  value is updated to  $R(S) = 1 - 1 = 0$ . The host  $S$  can not maintain any further outgoing links.

The last figure 7e illustrates the attaching of the host  $v_4$ . The nearest neighbour of  $v_4$  is  $S$ . But  $R(S) = 0$  and there are no other free potential neighbours with the same topological distance. The algorithm  $JOIN_R$  checks in this case all potential nearest neighbours  $v$  whether any hosts  $v_x$  with  $D(v, v_4) \leq D(v, v_x)$  is connected to  $v$ . In our case:

$$\begin{aligned}
 D(S, v_4) = 0 &\leq D(S, v_1) = 1 \\
 D(S, v_4) = 0 &\leq D(S, v_2) = 2 \\
 D(S, v_4) = 0 &\leq D(S, v_3) = 1
 \end{aligned}$$

So the algorithm  $JOIN_R$  inserts the host  $v_4$  between  $S$  and  $v_1$  and updates the  $R(v_4)$  value accordingly  $R(v_4) = 3 - 1 = 2$ .

Similar to the algorithm  $JOIN$  (fig. 6) we show that the routing in the constructed multicast tree  $T$  is optimally organised, i.e. that  $S(T)$  is minimal *and* that the algorithm solves the resource capacity problem  $R(T)$  as defined below. In order to do so we assign to each vertex  $v \in V$  a resource value  $R(v)$  that represents the maximum number of outgoing data links which can be served by the vertex. We call  $R(T)$  *solved* iff  $\forall v \in V : R(v) \leq R_{max}(v)$ . Admittedly the algorithm constructs a multicast tree with minimal  $S(T)$  value and solves the resource capacity problem  $R(T)$  with respect to a following precondition:

$$\forall v \in V : R_{max}(v) > 0 \tag{8}$$

**Theorem 3.** *The algorithm  $JOIN_R$  (fig. 6) constructs a tree  $T$  with the minimal  $S(T)$  value and solves the resource capacity problem  $R(T)$ .*

*Proof.* The algorithm  $JOIN_R$  consists of two parts, each one performing a loop on the potential nearest neighbours  $v$  of the host *new*.

The first part is reduced to the algorithm  $JOIN$  (fig. 2) and proved by induction (theorem 1). If the first loop detects a nearest neighbour, then it must have at least one free outgoing link to attach a new vertex. So  $S(T)$  is minimal and  $R(T)$  is solved.

The second loop is only executed if all potential nearest neighbours have no capacity. According the precondition 8 each vertex must be able to serve at least one outgoing link. It follows that one of the potential nearest neighbours must be connected to a vertex  $x$  with the topological distance  $D(v, x) \geq D(v, new)$  (fig. 8a). The vertex  $x$  is reconnected to the vertex *new* (fig. 8b).  $D(v, x) = D(new, x)$ , because  $v$  is one of the nearest neighbours of *new*. This step can be reduced to the algorithm  $JOIN$  (fig. 2) and proved by induction (theorem 1). So  $S(T)$  is minimal. According the precondition 8 the vertex *new* must be able to serve an outgoing link to  $x$  and  $R(T)$  is solved.  $\square$

In the next section we present an extension of the algorithm to reduce the end to end delay.

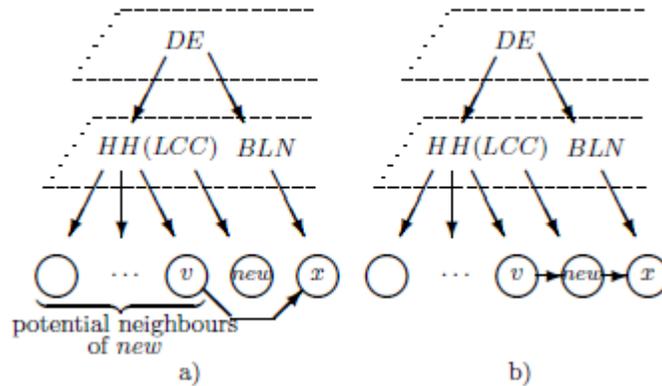


Figure 8. Proof of the second loop.

#### 4.4 End-to-End Delay

An important performance metric that is of concern for live media streaming overlays is the End-to-End Delay.

The End-to-End Delay ( $EED$ ) depends on the underlying network delay and the local delays at overlay peers due to queuing and processing. Formally we define the  $EED$  as a delay of the path  $p = \{v_0, \dots, v_n\}$  from the source host  $v_o$  to the end host  $v_n$ :

$$EED = \sum_{i=0}^{|p|} D(v_i, v_{i+1}) + |p| \quad (9)$$

We use the sum of the topological distances to represent the logical delay in the network and number of overlay hosts to represent the local delays at overlay peers. The algorithm  $JOIN_R$  constructs an overlay multicast tree  $T$  by choosing topologically close peers as neighbours. The  $S(T)$  value is minimal (theorem 3), so the delay in the network is minimal. To reduce total End-to-End Delay it is necessary to minimize the number of peers on the delivery path  $p$ . So the total End-to-End Delay depends on the multicast tree height. In order to keep the End-to-End Delay small we adapted the algorithm  $JOIN_R$  for construction a low stretched multicast tree as follows:

1. The loops on the potential nearest neighbours  $v$  of the host  $new$  are executed in the sorted order. The potential nearest neighbours are sorted by the End-to-End Delay to the source host according the equation (9).
2. The insert procedure (Fig. 9: lines 15-17) is modified. The host  $new$  is inserted between the host  $v$  and the host  $v_x$  with the lowest End-to-End Delay according to equation (9). And the outgoing links of  $v_x$  are reattached to  $new$  as long as  $R(new) > 0$ .

The pseudocode of the joining algorithm  $JOIN_{RE}$  is shown in figure 9 (we call  $v$  in  $LCC(new, n)$  iff  $LCC(new, v) = LCC(new, n)$ ). Since the basic structure of the algorithm  $JOIN_{RE}$  is equal to the structure of the algorithm  $JOIN_R$ , the algorithm satisfies the conditions of the theorem 1 and 3. For example in fig. 7c the potential nearest neighbours of the host  $v_2$  are  $S$  and  $v_1$ . But the host  $S$  has the lower  $EED$ -value, so the host  $v_2$  is attached to  $S$ . The host  $v_1$  is attached to  $S$  (fig. 7d), because  $S$  has the lowest  $EED$ -value and has enough resources.

In fig. 7e the host  $v_4$  can be inserted between three potential hosts  $v_3$ ,  $v_1$  and  $v_2$ . According to equation 9 the  $EED(S, v_3) = 2$ ,  $EED(S, v_1) = 2$  and  $EED(S, v_2) = 3$ . So the host  $v_4$  is inserted between  $S$  and  $v_1$ .

The modified algorithm  $JOIN_R$  keeps the End-to-End Delay small because the multicast tree height is logarithmic to the number of hosts. The total End-to-End Delay is  $O(\log N)$ .

---

Algorithm *JOINRE* (*new*,  $T_s$ )

---

```

1   $T_s = T_s \cup C(new)$ 
2  /*  $n$  is a potential neighbour of  $new$  */
3  for (all reachable hosts  $v$  in  $LCC(new, n)$ 
      in sorted order by EED ) {
4    if( $R(v) \neq 0$ ) {
5      attach  $new$  to  $v$ ;
6       $R(v) = R(v) - 1$ ;
7      return;
8    }
9  }
10 for (all reachable hosts  $v$  in  $LCC(new, n)$ 
      in sorted order by EED ) {
11  for (all hosts  $v_x$  connected to  $v$ ) {
12    if( $D(v, new) \leq D(v, v_x)$ ) {
13      insert  $new$  between  $v$  and  $v_x$ ;
14       $R(new) = R(new) - 1$ ;
15      while( $R(new) > 0$  or
             $v_x$  has outgoing links) {
16        reattach an outgoing link of  $v_x$  to  $new$ ;
17      }
18      return;
19    }
20  }
21 }

```

---

Figure 9. The pseudocode of the JOINRE algorithm

#### 4.5 Reconstruction algorithm

In order to support handling of peer connections we propose an algorithm to handle a host departure that may occur on purpose or by accident accidentally.

The pseudocode of the reconstruction algorithm is shown in figure 10. The algorithm deletes a host if it is a leaf. Otherwise it tries to reattach the outgoing links to the parent as long as it has enough resources. Finally the algorithm executes the algorithm *JOIN* for remaining hosts.

### 5. EXPERIMENTAL EVALUATIONS

In this section, we present simulation results for evaluation of the proposed approach. We have implemented a simulation to create overlay multicast trees according to our approach and as a balanced binary tree. The simulation parameters are: number of hosts and the resource capacity of each host  $R$ . The simulation calculates the routing costs in the tree  $S(T)$  according the definition in section 4.2, the height of the tree  $H(T)$  and the End-to-End Delay (EED) according to the equation 9 (section 4.4).

The tables 1 and 2 present the results of the comparison a multicast tree constructed by the *JOIN*-Algorithm with a randomly constructed balanced binary tree with the  $R(v) = 3$

---

**Algorithm RECONSTRUCT\_TREE ( $v_{del}$ )**

---

```

1  if ( $v_{del}$  is a leaf) {
2      delete  $v_{del}$ ;
3  } else {
4       $v_p$  = parent of  $v_{del}$ ;
5      for (each child  $v$  of  $v_{del}$  in sorted order) {
6          if( $R(v_p) > 0$ ) {
7              reattach  $v$  to  $v_p$ ;
8               $R(v_p) = R(v_p) - 1$ ;
9          }
10         else {
11             JOIN( $v, T_s$ );
12         }
13     }
14 }

```

---

Figure 10. The pseudocode of the reconstruction algorithm

**Table 1.** Comparison with  $R(v) = 3$ : JOIN - proposed algorithm, BBT - Balanced Binary Tree

# hosts	JOIN			BBT		
	$S(T)$	$H(T)$	$EDD$	$S(T)$	$H(T)$	$EDD$
10	13	3	4	17	4	10
100	103	5	7	180	10	20
500	503	6	8	850	14	28
1000	1003	7	9	1650	30	62

and  $R(v) = 2$  accordingly. The results show the total routing costs in the tree  $S(T)$  have

**Table 2.** Comparison with  $R(v) = 2$ : JOIN - proposed algorithm, BBT - Balanced Binary Tree

# hosts	JOIN			BBT		
	$S(T)$	$H(T)$	$EDD$	$S(T)$	$H(T)$	$EDD$
10	13	3	5	18	5	12
100	103	6	8	190	14	26
500	503	8	10	920	22	40
1000	1003	9	11	1890	44	78

the same value in the tree with  $R(v) = 2$  and  $R(v) = 3$ . The reason is that the algorithm always chooses the nearest neighbour and the routing costs are kept minimal. The End-to-End Delay depends on the height of the tree and the  $R(v)$  value respectively.

The chart 11 shows the End-to-End delay dependency graphically.

If the  $EDD$ -value of the tree generated by the algorithm *JOIN* increases only slightly, then the  $EDD$ -value of the binary balanced tree generated increases dramatically. In the real *P2P* overlay multicast tree each peer is able to serve different number of outgoing resources. The trend of  $EDD$  delay dependency will remain definitely similar.

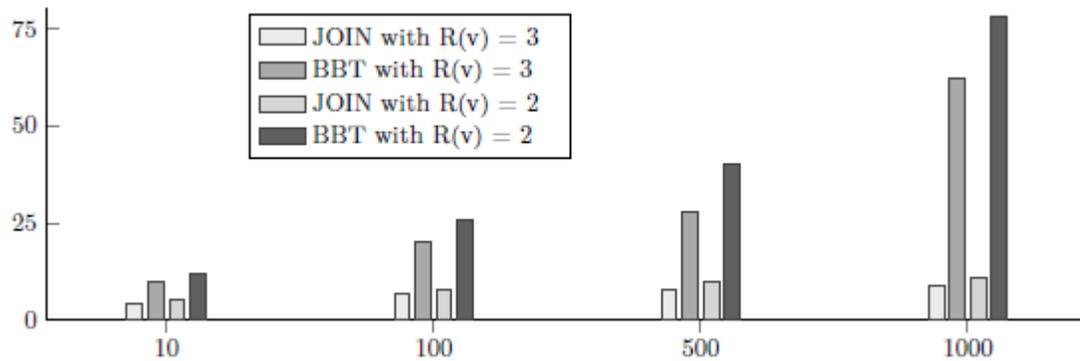


Figure 11. Comparison of results: JOIN - proposed algorithm, BBT - Balanced Binary Tree

## 6. CONCLUSION

In this paper we presented a novel multicast tree construction and maintenance approach based on the topological network coordinates of the end hosts. The algorithms presented in this paper were developed to achieve the following desirable properties:

- Minimal routing overhead in the underlying network
- Optimal resource management of the hosts
- Short end-to-end delay

We evaluated our approach theoretically and by using simulations. Compared to the randomly generated trees our approach improves significantly the performance metrics of a multicast overlay tree. Our future work will concentrate on implementing this approach in a real environment, collecting and analysing the performance data.

## REFERENCES

- [1] Abboud, O., Pussep, K., Kovacevic, A., Mohr, K., Kaune, S., Steinmetz, R., Enabling Resilient P2P Video Streaming, *Multimedia Systems*, Vol. 17, No. 3, p. 177-197, June 2011
- [2] Jurca, D., Chakareski, J., Wagner, J., Frossard, P., Enabling Adaptive Video Streaming in P2P Systems, *IEEE Communications Magazine*, p. 108-114, June 2007
- [3] Tran, D. A., Hua, K., Do, T., ZIGZAG: An Efficient Peer-to-Peer Scheme for Media Streaming, *Proc. of IEEE INFOCOM*, Vol.2, pp.1283-1292, 2003
- [4] Márk Jelasity and Ozalp Babaoglu, T-Man: Gossip-based overlay topology management, *3rd Int. Workshop on Engineering Self-Organising Applications (ESOA'05)*, Springer-Verlag, pp. 1-15, 2005
- [5] X. Liao, H. Jin, Y. Liu, L. M. Ni, D. Deng., AnySee: Peer-to-peer live streaming. In *Proceedings of IEEE International Conference on Computer Communications*, Barcelona, Spain, 2006.
- [6] Magharei, N., Rejaie, R., Yang G., Mesh or Multiple-Tree: A Comparative Study of Live P2P Streaming Approaches, *26th IEEE International Conference on Computer Communications-INFOCOM*, pp. 1424-1432, 2007.

- [7] H. Byun and M. Lee, HyPO: A Peer-to-Peer based Hybrid Overlay Structure, IEEE ICACT 2009, Feb. 2009.
- [8] Awiphan, S., Zhou Su, Katto, J., Two-layer Mesh/Tree Overlay Structure for Live Video Streaming in P2P Networks, proc. 7th IEEE Consumer Communications and Networking Conference (CCNC), pp. 1-5, 2010
- [9] F., Wang, Y., Xiong, and J., Liu, mTreebone: A Collaborative Tree-Mesh Overlay Network for Multicast Video Streaming, IEEE Transactions on Parallel and Distributed Systems, Vol. 21, No. 3, pp. 379-392, March 2010.
- [10] Donald Knuth, The Art of Computer Programming, Addison-Wesley, Vol. 3, 1973
- [11] Zhang, X.Y., Zhang, Q., Zhang, Z., Song, G., Zhu, W., A Construction of Locality-aware Overlay Network: mOverlay and its Performance, IEEE Journal on Selected Areas in Communications, pp. 18-28, 2004
- [12] Banerjee, S., Bhattacharjee, B. Kommareddy, C., Scalable application layer multicast, Proc. ACM SIGCOMM Conf., ACM Press, New York, 2002.
- [13] Abboud, O., Kovacevic, A., Graffi, K., Pussep, K., Steinmetz, R., Underlay Awareness in P2P Systems: Techniques and Challenges, IEEE International Parallel and Distributed Processing Symposium, 2009
- [14] Xuping Tu, Hai Jin, Xiaofei Liao, and Jiannong Cao, Nearcast: A locality-aware P2P live streaming approach for distance education. ACM Transactions on Internet Technology, Vol. 8 - Issue 2, 2008
- [15] T. S. Eugene Ng, Hui Zhang, Predicting internet network distance with coordinates-based approaches. Proc. of IEEE INFOCOM, New York, Vol. 1, pp. 170–179, 2001
- [16] James A. Muir and Paul C. Van Oorschot, Internet geolocation: Evasion and counterevasion, Journal ACM Computing Surveys, Vol. 42 - Issue 1, No. 4, December 2009
- [17] Editor: Andrei Popescu, Geolocation API Specification, W3C, 22 December 2008
- [18] Editor: Philip Olson, PHP Manual - Geo IP Location, The PHP Documentation Group, 2014, <http://php.net/manual/en/book.geoip.php>
- [19] Chao Dai, Yong Jiang, Shu-Tao Xia, Hai-Tao Zheng, and Laizhong Cui. A traffic localization strategy for peer-to-peer live streaming. In 2013 IEEE Symposium on Computers and Communications, ISCC 2013, Split, Croatia, 7-10 July, 2013, pages 495–501, 2013.
- [20] Paul Francis, Sugih Jamin, Vern Paxson, Lixia Zhang, Daniel F. Gryniewicz, and Yixin Jin. An architecture for a global internet host distance estimation service, Proceedings of IEEE INFOCOM, 1999.
- [21] Ethan Katz-bassett, John P. John, Arvind Krishnamurthy, David Wetherall, Thomas Anderson, and Yatin Chawathe. Towards IP Geolocation Using Delay and Topology Measurements, IMC, 2006

*INTENTIONAL BLANK*

# JPL : IMPLEMENTATION OF A PROLOG SYSTEM SUPPORTING INCREMENTAL TABULATION

Taher Ali<sup>1</sup>, Ziad Najem<sup>2</sup>, and Mohd Sapiyan<sup>1</sup>

<sup>1</sup>Department of Computer Science, Gulf University for Science and Technology, Kuwait

ali.t@gust.edu.kw, sapiyan.m@gust.edu.kw

<sup>2</sup>Department of Computer Science, Kuwait University, Kuwait

najem@cs.ku.edu.kw

## **ABSTRACT**

*The incremental evaluation of tabled Prolog programs allows to maintain the correctness and completeness of the tabled answers under the dynamic state. This paper presents JPL implementation details. JPL is an approach to support incremental tabulation for logic programs under non-monotonic logic. The main idea is to cache the proof generated by the deductive inference engine rather than the end results. In order to be able to efficiently maintain the proof to be updated, the proof structure is converted into a justification-based truth-maintenance (JTMS) network.*

## **KEYWORDS**

*Applications of justification-based truth maintenance systems, Belief revision systems, Truth maintenance systems, Justification-based truth maintenance systems, Incremental evaluation of tabled Prolog, Incremental tabulation for Prolog queries, Tabulation for logic programs, Memoing for logic programs.*

## **1. INTRODUCTION**

Prolog is a logic programming language associated with artificial intelligence and computational linguistics [1, 2, 3]. Tabled extension for Logic Programming (TLP) [4, 5, 6] evaluation enhances the performance of the Prolog query engine and allows the termination of certain computations that do not terminate under the normal Prolog query evaluation. The incremental evaluation [7, 8] of tabled Prolog programs allows to maintain the correctness (soundness) and completeness of the tabled answers under the dynamic state. This evaluation strategy allows the system to update the tabled answers when the set of facts and/or rules participated in generating the answers of a certain query are either retracted from or asserted to the Prolog program. JPL [7, 9] presented an approach of maintaining one consolidate system to cache the query answers under the non-monotonic logic. It uses the justification-based truth-maintenance system to support the incremental evaluation of tabled Prolog Programs. In this paper we will focus on the system

implementation. The system implementation must take into consideration the following performance factors:

1. Minimizing the overhead of caching the query proof structure.
2. Minimizing the time and space needed to maintain soundness and completeness of the cached proof structure.

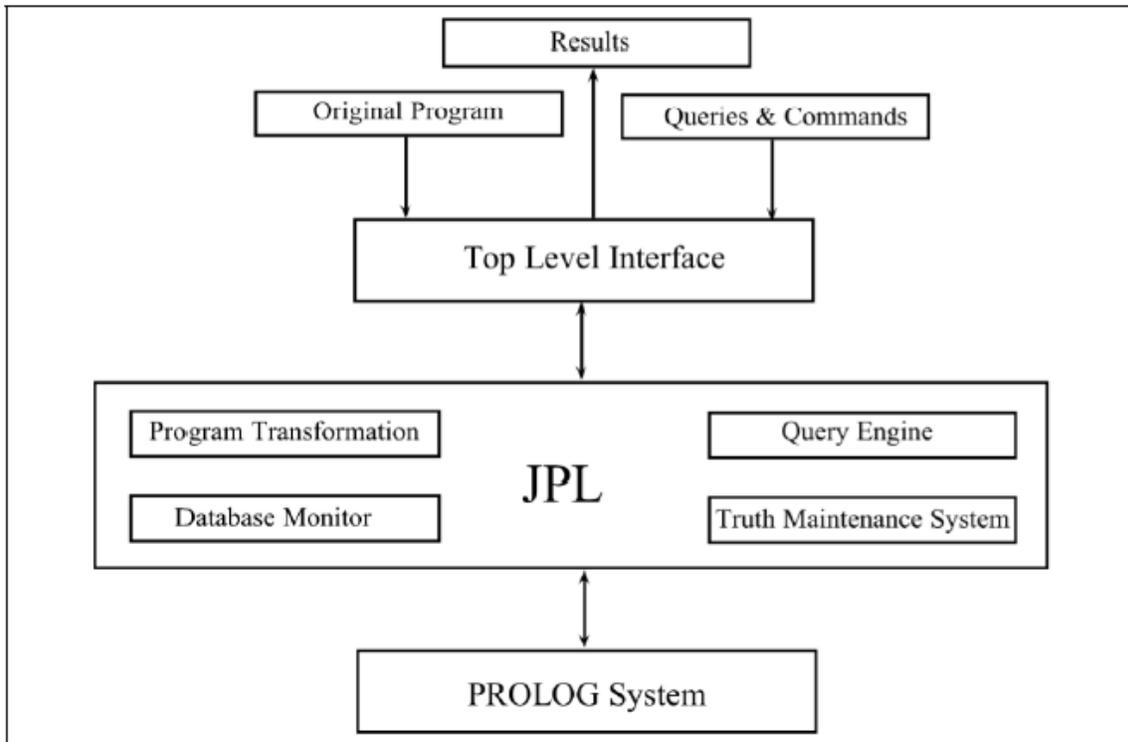


Figure 1: An overview of JPL

Another factor that we have to take into consideration regarding the system implementation is the portability of integrating JPL with more than one PROLOG inference engine.

## 2. RELATED WORK

Tabling is an implementation technique where answers for subcomputations are stored and then reused when a repeated computation appears. There are two approaches to integrate tabling support into existing PROLOG systems. The first approach is to modify and extend the low-level engine. This is the common approach used by most of systems to support tabled evaluation. The approach modifies and extends the low-level engine [10]. The advantage is the run-time efficiency, however, the drawback is that it is not efficiently portable [11] to other Prolog systems because the engine level modifications are slightly more complex and time consuming. The second approach to incorporate tabled evaluation into existing Prolog systems is to apply the source level transformations to a tabled program, and then use external tabling primitives to provide direct control over the search strategy. This idea was first explored by Fan and Dietrich

[12] and later used by Rocha, Silva and Lopes [13] to implement tabled PROLOG systems. The main advantage of this approach is the portability to apply the approach to different PROLOG systems. The drawback of course is the efficiency, since the implementation is not at a low level.

---

#### Algorithm 1 Program Transformation Algorithm

---

Input:  $R$ , original Program rules list

Output:  $R_t$ , transferred rules List

```

assign each rule in R a unique id
for each rule Ri in R
    h = head of Ri, b = body of Ri;
    ht= h with extra unique var argument Jc;
    bt = b+", ";
    bt = bt + "append(["+id(Ri)+"],[";
    bt = bt + "["+all non-negated terms in b+"]);";
    bt = bt + "["+all negated terms in b+"],["+h+"]);";
    bt = bt + "],"+ Jc + " )";
    Rti = ht + ":-" + bt + ".";

```

---

### 3. JPL IMPLEMENTATION APPROACH

JPL implementation is based on applying the source level transformations to a tabled program. This will allow to incorporate the idea of incremental tabulation into different PROLOG systems using the same general framework. In order to integrate JPL with PROLOG inference engine we are using INTERPROLOG. INTERPROLOG [14] is an PROLOG-JAVA application programming interface that supports multiple PROLOG systems through the same API. The current version (2.1.2a at the time of implementing this approach) of INTERPROLOG supports three PROLOG implementations (XSB [15], SWI-PROLOG [16] and YAP-PROLOG [17]) on different platforms (Windows, Linux and Mac OS X). It promotes coarse-grained integration between logic and object-oriented layers, by providing the ability to bidirectionally map any class data structure to a PROLOG term. This basic idea of INTERPROLOG is allowing PROLOG to call any JAVA method, and for JAVA to invoke PROLOG goals. This is achieved by using a communication layer to pass object/term data among both processes. INTERPROLOG is used in the implementation of JPL to provide an interface from JPL to all PROLOG inference engines supported by INTERPROLOG.

### 4. PROGRAM TRANSFORMATION

PROLOG rules are written in a standard form known as Horn clauses. A Horn clause statement has the form:

$$H : -B_1, B_2, \dots, B_n.$$

where  $H$  is a first-order atom and each  $B_i$  is a first-order literal.  $H$  is called the head of the clause and  $B_1, B_2, \dots, B_n$  is the body of the clause. The semantic of this statement is that when  $B_i$  all are true, we can deduce that  $H$  is true as well. The above clause can be read " $H$ , if  $B_1, B_2, \dots, B_n$ ". This basic concept of logic programming is used by JPL to do the program transformation. When a PROLOG programmer executes the command `consult/1` to read a PROLOG source file through the JPL's top level interface, JPL executes the program transformation method to convert the

PROLOG predicates (rules) into a format that allows the system later to table the query answers as a JTMS network according to the mechanisms described in the previous chapter.

JPL applies program transformation only on clauses that are selected by the PROLOG programmer to be as incremental tabled predicates. Algorithm 1 illustrates the program transformation process. The process starts with assigning each rule in the original program a unique ID.

```

%Original Tabled Predicates
connected(X,Y) :- edge(X,Y). %r1
connected(X,Y) :- edge(X,Z), connected(Z,Y). %r2

%Transformed Predicates by JPL
connected(X,Y,J0) :- edge(X,Y), append([r1], [[edge(X,Y)], [], connected(X,Y)], J0).
connected(X,Y,J0) :-
-edge(X,Z), connected(Z,Y,J1), append([r2], [[edge(X,Z), connected(Z,Y)], [], connected(X,Y)], J0).

```

Figure 2: Original tabled and transformed Predicates by JPL for the translative closure program of the directed edge relationship.

For example, The first rule in the translative closure program of Figure 2 is uniquely identified as r1, while the other rule is defined as r2. The next step in the algorithm is to convert each rule in the original program into a format that allows the query engine later to link the facts which are participated as antecedents to produce the consequence (answer) of the query. This is achieved by applying two major changes in each program rule:

1. Add an extra unique var argument to the rule head, this var points to a list which contains all the necessary information needed by JPL to build the JTMS network of any query answer generated from this rule. The list is to be composed of the following items:
  - (a) The rule ID.
  - (b) List of facts (antecedents), coming from the rule body, that must be true (IN) to support the rule consequence.
  - (c) List of facts (antecedents), that must be false (OUT) to support the rule consequence. This case is used when the body contains negated terms.
  - (d) The consequence of the rule, which is basically the head of the rule.
2. Add an extra term to the rule body that generates the above list and stores it in the extra var added to rule head.

Once the rule is in its final format using the above conversion method, a TMS node is created for the rule and it is linked to the TMS node of the original rule head. This will help the query engine later to identify all rules related to a certain query entered by the user. Figure 2 shows how a *connected/2* tabled predicate that defines the translative closure program of the directed edge relationship is transformed to support JPL's tabulation strategy. Each rule head, in Figure 2, is converted from *connected(X,Y)* to *connected(X,Y,J0)* by adding an extra var argument J0. In the same manner each rule body is changed by appending the required code that originates the list in the var J0. For example the second rule body is changed from :

*edge(X,Z), connected(Z,Y)*

to:

*edge(X,Z); connected(Z,Y,J1); append([r2]; [[edge(X,Z), connected(Z,Y)]]; []; connected(X,Y)]; J0):*

Both of these two rules are linked to the TMS node of the original rules head, i.e. *connected(X,Y)*.

## 5. QUERY ENGINE

---

### Algorithm 2 Query Engine

---

Input: *Q*, a PROLOG query.

Output: List of query answers.

```

queryTmsNode = tmsNode associated with Q
if ( queryTmsNode != full ){
    tn = getGrandParent( Query );
    gt = getRules( tn );
    while ( gt != null ){
        justifications=executePrologQuery( gt.element.rhs );
        installJustifications( justifications );
    }
    justifications=executePrologQuery( Query );
    installJustifications( justifications );
}
}
showQuerySolutions( queryTmsNode );

```

---

Query engine of JPL responds to end user queries through the system top interface. The necessary methods needed to implement the logic of the query engine are part of the JTMS class. There are two scenarios for the query execution module that are described in Algorithm 2. The first scenario checks if the TMS node associated with the user query already exists in the JTMS network and is marked as fully proved. When this condition is true the query engine simply shows all the valid(IN) answers of the query from the JTMS network linked to the query's TMS node. The second scenario deals with the case when the query is executed for the first time. JPL deals with this case as a three step process:

1. The system locates the general TMS node associated with the query. The desired TMS node is used to locate the set of PROLOG rules associated with the query. If the set is not empty, the query engine requests the PROLOG abstract engine attached to it to execute the right hand side of each rule. The answers (justifications) coming from the PROLOG abstract engine execution are installed as justifications in the JTMS network.
2. The query engine tries to find more solutions for the query that might come from the database of facts which were not discovered in the first step. If such answers exist, then justifications related to these answers are also installed in the JTMS network.

3. After the previous two steps, the query answers are ready. The query engine simply shows all the answers of the query from the JTMS network linked to the query's TMS node.

### 5.1 Installing the Justifications

As seen in Section 5, the PROLOG abstract engine, attached to JPL, executes PROLOG queries requested by the query engine. The results of these queries must be installed as justifications in the JTMS network associated with the query. Figure 3 represents the list of answers (b) returned by the PROLOG abstract engine for the query `connected(X,Y)` with respect to the translative closure PROLOG program (a). The list of results are in a format that allows the system to convert them easily into justifications. The list item contains the rule participated as antecedent in generating the deduction. The set of facts participated as antecedents in generating the deduction.

```
connected(X,Y) : -edge(X,Y). %r1
connected(X,Y) : -edge(X,Z),connected(Z,Y). %r2
edge(a,b). edge(a,c). edge(b,d). edge(c,d). edge(d,e). %f1..f5
```

(a)

```
[r2,[edge(a,b),connected(b,d)],[],connected(a,d)]
[r2,[edge(a,c),connected(c,d)],[],connected(a,d)]
[r2,[edge(a,b),connected(b,e)],[],connected(a,e)]
[r2,[edge(a,c),connected(c,e)],[],connected(a,e)]
[r1,[edge(a,b)],[],connected(a,b)]
[r1,[edge(a,c)],[],connected(a,c)]
[r2,[edge(b,d),connected(d,e)],[],connected(b,e)]
[r1,[edge(b,d)],[],connected(b,d)]
[r2,[edge(c,d),connected(d,e)],[],connected(c,e)]
[r1,[edge(c,d)],[],connected(c,d)], [r1,[edge(d,e)],[],connected(d,e)]
```

(b)

Figure 3 : List of answers(b) returned by the PROLOG abstract engine for the query `connected(X,Y)` with respect to the translative closure PROLOG program(a).

The answer itself which represents the consequence of the deduction. Below are the details of each list item:

1. The rule ID that is behind the generation of this result.
2. Set of PROLOG items that must be IN to support the consequence of the result. This case is related to non-negated subgoals in the right hand side of the rule.
3. Set of PROLOG items that must be OUT to support the consequence of the result. This case is related to negated subgoals in the right hand side of the rule.
4. The consequence of this result.

These four items are one to one mappings to the data attributes for the Justification class. The class contains five attributes. The first attribute is used as a flag to indicate whether the

justification is active or not. The rest of the attributes are used to store the above information for the justification.

Going back to the example of Figure 3(b), the first answer:

$$[r2; [edge(a,b); connected(b,d)]; []; connected(a,d)]$$

for the query  $connected(X,Y)$  states the following:

1. The rule  $r2$  is in the antecedent list of the answer  $connected(a,d)$ .
2. The atoms  $edge(a,b)$  and  $connected(b,d)$  must be IN to support the consequence of the the answer  $connected(a,d)$ .
3. None of the atoms must be OUT to support the consequence of the answer since there are no negated subgoals in the program of Figure 3(a).
4. The consequence of this answer is the Prolog atom  $connected(a,d)$ .

For each of the answers of Figure 3(b), JPL installs a justification for it and makes the justification active. When the data related to any justification is changed, the system maintains the consistency of the justification according to the changes in the set of rules/-facts related to the justification. An important point that must be highlighted is that each justification element is stored as a TMS node. The details of the TMS node data structure is given in next section.

## 6. TMS NODES

A TMS node is the basic data structure used in JPL to cache the proof structure of a PROLOG query. The set of TMS nodes linked together through justifications represents the JTMS network for the query. The TMS node is a complicated data structure, it must store all the information that allows the system to maintain the consistency and completeness of the cached proof structure for a previously proven query. The following subsections describe the main most important attributes and operations for this crucial data structure.

### 6.1 Attributes

#### Label

The label attribute stores the current status of the node. At any time, the TMS node is labeled one of the following:

1. IN or OUT

One of these two labels are used to represent the status of the TMS node when the TMS node that is pointing to a PROLOG predicate, i.e. a fact or a rule. When the system believes that the PROLOG predicate is true or active, then the label of the TMS node is IN. The TMS node label is OUT when the predicate is false or inactive.

## 2. Full, Partial or New

One of these three labels are used when the TMS node is pointing to a PROLOG query. The label Full indicates that the query associated with the TMS node is fully proved and when the user re-evaluates the query no additional work would be required from the PROLOG inference engine attached to JPL in generating the answers of the query. The label Partial means that the query attached to the TMS node is partially proved. If a partially proven query is going to be evaluated then the system must complete the query evaluation before returning its answers. The New label is used when the query is proved for the first time and a new TMS node is created for the query.

### Type

This attribute stores the type of the TMS node. JPL uses three types of TMS nodes. The type of the TMS node is either a PROLOG fact, rule or a query.

### Node

The Node attribute points to the actual PROLOG term.

### Support

This object represents the set of justifications that support the PROLOG fact attached to this TMS node.

---

#### Algorithm 3 Set label operation's algorithm for the TMS node

---

Input: *newLabel*, the new label

Output: *void*

```

if (this.label != newLabel) {
    boolean active = false;
    if ( this.type == Fact && newLabel == OUT ){
        active = this.findAnotherSupport();
        if (!active) {
            this.label = x;
            if (this.type == 'Fact' || this.type == 'Rule')
                this.propagateInnessOutness(myJtms);
        }
    }
}

```

---

### In-List

The set of justifications where the PROLOG predicate, associated with this TMS node, is participating as an IN(*true*) antecedent in the justification.

### Out-List

The set of justifications where the PROLOG atom, associated with the this TMS node, is participating as an OUT(*false*) antecedent in the justification.

## Related Queries

This set points to the list of other cached queries, in the system, which are effected whenever there is a change in the proof structure of the PROLOG query attached to this TMS node.

## Rules

If the TMS node is associated with a PROLOG query where the query term matches the head of certain rules in the PROLOG program, then this attribute points to the query related rules from the program. Note that some of the above attributes might have a Null value depending on the type of the TMS node. For example, if the TMS node is associated with a PROLOG query, then the values of support, in-list and out-list is Null since they are used with TMS nodes that point to a PROLOG fact.

## 6.2 Operations

### Set Label

The "Set Label" is the most critical operation of JPL. The operation is maintaining the consistency and the completeness for the queries cached proof structure. Algorithm 3 describes the set label process for a TMS node. The algorithm starts with checking if the new label is different from the current label of the TMS node. This check is required to avoid any redundant computations when there is no change in the label. The next step in the algorithm is to check if the label of the TMS node is being changed from IN to OUT and if that particular node is pointing to a PROLOG fact. If this is true, then the system tries to find an active justification that can support this fact, and hence, avoid changing its label from IN to OUT. If no such active justification is found, or the TMS node type is not a fact, or the new label is being changed from OUT to IN, then the system changes the label of the TMS node to its new value. Once the label value is changed, JPL propagates the effect of this change throughout the JTMS network by calling the propagateInnessOutness method.

---

**Algorithm 4** Propagating the change of a TMS node label throughout the JTMS network.

---

Input: *this*, the node with changed label.

Output: *void*

```

for each justification in this.inList {
    if (this.label == IN)
        try to make the justification active;
    else
        make the justification inactive;
}
for each justification in this.outList {
    if (this.label == OUT)
        try to make the justification active;
    else
        make the justification inactive;
}

```

---

### Propagate Inness/Outness of Existing Facts/Rules

The objective of this operation is to maintain the consistency of the JTMS network whenever there is a change in the label of a TMS node. The algorithm for propagating the change of a TMS node label throughout the JTMS network is explicated in Algorithm 4. The logic of the algorithm is straightforward. For each justification in the in-list of the current TMS node, the system tries to set the justification active when the label of TMS node is IN; otherwise the justification is set as inactive. Also, for each justification in the out-list of the current TMS node, the system tries to set the justification active when the label of the TMS node is OUT, otherwise the justification is set as inactive.

- Activating a justification

To activate a justification, the system must ensure that all antecedents of the In-List are labeled as IN, and all antecedents of the Out-List are labeled as OUT. If these two conditions are satisfied, then the justification's consequent TMS node label is changed to IN, and the justification is marked as active.

- Inactivating a justification

To mark a justification as inactive, the system changes the justification's consequent TMS node label to OUT. Then the justification is marked as inactive.

### Propagate the Asserting of a New Fact

This method is called when a new PROLOG fact is asserted through the JPL interface. Figure 5 outlines the algorithm for propagating the effect of inserting a new PROLOG fact. The system locates the TMS query nodes that might get affected from the insertion of the new fact. For each such query, if the query has been fully proved previously, the algorithm locates the set of rules attached to the query's TMS node. Every rule is unified with the new fact, then the system executes the partial PROLOG query (which is coming from the right hand side of the rule) to update the cached proof structure.

---

**Algorithm 5** Propagating the effect of asserting a new PROLOG fact.

---

Input: *this*, the new PROLOG fact's TMS node.

Output: *void*

```

RQ= this.relatedQueries;
for (each Query in RQ) {
    if (Query.label == Full) {
        Rules= Query.getRules();
        for (each Rule in Rules) {
            Query.executePartialQuery(unify(Rule, this));
        }
    }
}

```

---

---

**Algorithm 6** Propagating the effect of asserting a new PROLOG rule.

---

Input: *this*, the new PROLOG rule's TMS node.Output: *void*

```

ruleQueryNode=tms.find(this.head);
if (ruleQuery != null) {
    if (ruleQuery.getLabel() == Full) {
        Jtms.executePrologRule(head, this);
    }
}

```

---

**Propagate the Asserting of a New Rule**

This method is called when a new PROLOG rule is asserted through the JPL interface. Figure 6 outlines the algorithm for propagating the effect of inserting a new PROLOG rule. The process starts with locating the TMS query node that matches the new rule head. If such query node exists and it is marked as fully proved, then the system requests the PROLOG inference engine attached to JPL to execute the right hand side of the new rule to update the cached proof structure. This is achieved by invoking the JTMS method *executePrologRule*.

**7. The JTMS CLASS**

The JTMS class is the main component of JPL. It interacts with the system's top interface (JPL class) to provide the desired functionality (see Figure 2). The main data component and operations of this class are described briefly below.

**7.1 Attributes****PROLOG Engine**

The PROLOG engine points to the PROLOG inference engine. The integration between the PROLOG inference engine and JPL is done by using InterProlog package. See Section 3 for more details about InterProlog.

**TMS Object**

The object points to an instance of the TMS class. See Section 8 for more details about this component.

---

**Algorithm 7** Executing a PROLOG Query

---

Input: *query*, the PROLOG query(goal).Output: *void*,

```

solutionVars=engine.deterministicGoal(query);
for (each solution in solutionVars) {
    this.installJustifications(solution);
}

```

---

---

**Algorithm 8** Handling the the assertion of a fact/rule to the PROLOG program.

---

Input: *term*, the asserted PROLOG fact or rule.Output: *void*,

```

engine.deterministicGoal(assert(term));
TmsNode t1 = myTms.findNode(term);
if (t != Null) {
    t1.setLabel(IN)
} else {
    if (term == Fact) {
        Tms newNode = myTms.addNode(term);
        newNode.propagateNewAtom();
    } else {
        Term newRule = convert(term);
        TmsNode newNode = myTms.addNode(newRule);
        newNode.propagateNewRule();
    }
}

```

---

**Justifications Table**

The justification object of JTMS class points to the list of all current justifications in the system. Justifications are maintained as a hash table to allow fast retrieval of information related to them. See Section 5.1 for more details about how the justifications are installed and maintained by JPL.

**7.2 Operations****Executing PROLOG Queries**

This method allows the JTMS class to execute the Prolog queries with the help of the Prolog inference engine object. Usually this method is called by the query engine when JPL's end user requests to execute the query for the first time. Later, the method is invoked by the TMS node objects to maintain the correctness and completeness of the cached proof structure for the same query. Algorithm 7 outlines the execution of a Prolog query. This is a two step process. First the incoming query is executed by invoking the method *deterministicGoal* which returns the set of solutions for this particular query. In the next step, the algorithm takes each solution returned by the Prolog inference engine and calls the method *installJustifications* to install the justification in the JTMS network of JPL.

---

**Algorithm 9** Handling the the retraction of a fact/rule from the PROLOG program.

---

Input: *term*, the retracted PROLOG fact or rule.Output: *void*,

```

engine.deterministicGoal(retract(term));
if (t != Null) {
    TmsNode t1 = myTms.findNode(term);
    t1.setLabel(OUT)
}

```

---

## **jAssert**

The method *jAssert* is invoked when the end user initiates the PROLOG command `assert/1`, through JPL's user interface to insert a fact/rule into a PROLOG program. Algorithm 8 describes the steps to handle the assertion of a fact/rule to the PROLOG program. The first step in the algorithm is to inform the PROLOG engine about this change in the set of facts/rules related to the associated program. This is achieved by executing the `assert` command by the PROLOG engine. In the next step, the algorithm tries to locate the TMS node associated with the asserted fact/rule. If the TMS node already exists in the database of TMS nodes then the method `setLabel` (See Section 6.2) is invoked to propagate the effect of this assertion through out the JTMS network.

If the TMS node, associated with the asserted fact/rule, does not exist in the database of TMS nodes, then this case is related to the insertion of a new PROLOG fact/rule to the PROLOG program which was not presented at the program consult time. The system handles the insertion of new PROLOG predicates according to the following scenarios:

1. Asserting a new fact

A new TMS node is created for the new fact. The algorithm 8 then invokes the method *propagateNewAtom* (See Section 6.2) to update the JTMS network in response to this change in data.

2. Asserting a new rule

A new TMS node is created for the new rule. The rule is converted according to the format described in Section 4. Then the algorithm invokes the method *propagateNewRule* (See Section 6.2) to update the JTMS network in response to this change in the set of rules.

## **jRetract**

The inverse logic of *jAssert*. This method is invoked when the end-user initiates the Prolog command `retract=1`, through JPL's user interface, to remove a fact/rule from the Prolog program. Algorithm 9 shows the three required steps to handle the retraction of a fact/rule from the Prolog program. In the first step, the system informs the Prolog engine about this change in the database of facts/rules by executing the `retract` command on the Prolog engine. Then the algorithm tries to locate the TMS node associated with the asserted fact/rule. If the TMS node already exist in the database of TMS nodes, then the algorithm invokes the `setLabel` method to propagate the effect of this retraction throughout the JTMS network. If the TMS node associated with the retracted fact/rule does not exist in the database of TMS nodes, then no action is required because the retracted fact/rule is not part of the current JTMS network.

## **8. TMS CLASS**

The TMS class is used in JPL to link all TMS nodes together in a parent/child relationship. Basically, the TMS class stores and links the TMS Nodes using the Graph data structure.

---

**Algorithm 10** Adding a PROLOG term to the TMS Network.

---

Input: term, the PROLOG term; type, the type of the PROLOG term; label, the label of the PROLOG term.

Output: t, the TMS node associated with the PROLOG term.

```

t1 = new TmsNode(node, label, type);
this.addVertex(t1, null);
TmsNode t2=this.addNode(t1.getGrandParent(), 'I', New);
this.addEdge(t2, t1);
switch (nodeType) {
    case Fact:
        t1.propagateNewAtom();
        break;
    case Rule:
        t1.propagateNewRule();
        break;
}
return t1;

```

---

In the rest of this section, a brief description is given about the main components of the TMS class.

**Attributes**

There are two main attributes in this class. The first attribute is the *hashTable* that keeps the track of Prolog terms (facts, rules, and queries) with their associated TMS nodes in the *HashTable* data structure to allow fast retrieval of the terms whenever required by other system components. The other attribute of the TMS class is the *adjhash* which represents the adjacency list of each TMS node. The adjacency list of each TMS node links it to other TMS nodes in the JTMS network of JPL.

**Operations**

The operations of the TMS node are the typical ones of any Graph or *HashTable* abstract data type. Mainly the most important operations are:

**Add a TMS node**

The algorithm for adding a PROLOG term to the TMS Network is presented in 10. The algorithm starts with creating a new TMS node for the Prolog term. The newly created TMS node is added to the Graph as a vertex. Then the algorithm links the TMS node to its most general term. Once this setup is done, the algorithm propagates the effect of this new Prolog term throughout the JTMS network.

**Find a TMS Node**

Given a Prolog term (fact, rule, or query), the find operation returns a pointer to the TMS node object associated with the Prolog Term. If such TMS node does not exist in the *HashTable*, then the method returns null.

## 9. CONCLUSION

This paper presented JPL implementation details. The main aim of this research is to develop a tabled Prolog system that is capable of caching and maintaining the correct answers of the query under the non-monotonic logic in an efficient way. The design try to avoid, as mush as possible, the limitations and disadvantages in the existing approaches to support incremental evaluation of tabled PROLOG programs. To judge that our design meets it's objectives, a comprehensive performance analysis of JPL is required. This will be the target of future work.

## REFERENCES

- [1] W F. Clocksin and Christopher S. Mellish. Programming in Prolog (2nd ed.). Springer-Verlag New York, Inc., New York, NY, USA, 1984.
- [2] Michael A. Covington. Natural Language Processing for Prolog Programmers. Prentice-Hall, Englewood Cliffs, NJ, 1994.
- [3] Ivan Bratko. Prolog Programming for Articial Intelligence. Pearson Addison-Wesley, Harlow, England, 3 edition, 2000.
- [4] Hisao Tamaki and Taisuke Sato. Old resolution with tabulation. In Proceedings on Third international conference on logic programming, pages 84 98, New York, NY, USA, 1986.Springer-Verlag New York, Inc.
- [5] David Scott Warren. Memoing for logic programs. Commun. ACM, 35(3):93 111, 1992.
- [6] Terrance Swift. Principles, practice, and applications of tabled logic programming. SIGSOFT Softw. Eng. Notes, 25(1):87 88, January 2000.
- [7] Taher Ali, Ziad Najem, and Mohd Sapiyan. Query proof structure caching for incremental evaluation of tabled prolog programs. International Journal of Computer Science and Information Technology, 6(1):311, Feb. 2014.
- [8] Diptikalyan Saha and C. R. Ramakrishnan. Incremental evaluation of tabled logic programs. In ICLP, pages 392 406, 2003.
- [9] Taher Ali, Ziad Najem, and Mohd Sapiyan. A belief revision system for logic programs. Computer Science and Information Technology, 4(9):227 231, Sep. 2014.
- [10] Konstantinos Sagonas, Terrance Swift, and David S. Warren. Xsb: An overview of its use and implementation. SUNY Stony Brook, pages 11794 4400, 1993.
- [11] Jan Wielemaker and Vítor Santos Costa. Portability of prolog programs: theory and casestudies. CoRR, abs/1009.3796, 2010.
- [12] Changquan Fan and Suzanne Wagner Dietrich. Extension table built-ins for prolog. Softw. Pract. Exper., 22(7):573 597, July 1992.
- [13] R. Rocha, C. Silva, and R. Lopes. Implementation of Suspension-Based Tabling in Prolog using External Primitives. In J. Neves, M. Santos, and J. Machado, editors, Local Proceedings of the 13th Portuguese Conference on Articial Intelligence, EPIA'2007, pages 11 22, Guimarães, Portugal, December 2007.
- [14] Miguel Calejo. Interprolog: Towards a declarative embedding of logic programming in java. In José Júlio Alferes and João Alexandre Leite, editors, JELIA, volume 3229 of Lecture Notes in Computer Science, pages 714 717. Springer, 2004.
- [15] Terrance Swift and David Scott Warren. Xsb: Extending prolog with tabled logic programming. TPLP, 12(1-2):157 187, 2012.
- [16] Jan Wielemaker, Tom Schrijvers, Markus Triska, and Torbjörn Lager. SWI-Prolog. Theory and Practice of Logic Programming, 12(1-2):67 96, 2012.
- [17] Vítor Santos Costa, Ricardo Rocha, and Luís Damas. The yap prolog system. TPLP, 12(1-2):5 34, 2012.

**AUTHORS**

**Taher M. Ali** received his B.Sc. and M.Sc. from Kuwait University and PhD from University of Malaysia. Dr. Taher has over 15 years of experience in higher education sector. The experience is divided between academia as faculty member and IT related jobs. He is currently working as an Assistant Professor of Computer Science in Gulf University for Science and Technology (ABET and ACCSB accredited), and also serving as a head of computer science department since Fall 2015. He has also served the university Chief Information Officer (CTO) between 2009-2015. During his time as CTO, he has stabilized the IT department infrastructure, services, policies and procedures. Under Dr. Taher leadership, the IT department at GUST university is fully maintained by in-house team with high utilization of open source technologies and effective integration of different systems such as: student information, learning management, attendance, security, access, web, library, work flow, faculty information, and other sub systems required to manage the daily business activities of students/faculty/staff. In 2013, Dr. Taher has received The Kuwait Electronic Award for Enriching e-Content under category of Under Category of Electronic Sciences.



**Ziad H. Najem** received his BSc from Kuwait University and Ms and PhD from University of Illinois at Urbana-Champaign. Prior to joining the Department of Computer Science at Kuwait University in 1999, Dr.Najem worked as a Scientific Researcher at Kuwait Institute for Scientific Research.



**Mohd Sapiyan Baba** is currently a Professor of Computer Science in Gulf University for Science and Technology, Kuwait. He was a lecturer in University of Malaya for more than 30 years, teaching Mathematics and Computer Science courses, and supervised numerous students for their research projects at undergraduate and postgraduate levels. His main research interest is in field of Artificial Intelligence (AI), in particular, the application of AI in Education



## AUTHOR INDEX

- Abdelmalek AMINE* 13  
*Ali Gholami* 229  
*Allel Hadjali* 97  
*Aouatif Amine* 87  
*Ashish Shrivastava* 53  
*Aymen Gammoudi* 97  
*Bin Yang* 77  
*Bouchra Nassih* 87  
*Boutheina Ben Yaghlane* 97  
*Christian Kreiner* 183  
*Cordero, P* 159  
*Daniel Homm* 241  
*Djamel Eddine Saidouni* 297  
*Enciso, M* 159  
*Erwin Laure* 229  
*Eugen Brenner* 183  
*Fatima KABLI* 13  
*Florian Bock* 241  
*Georg Macher* 183  
*Gholamreza Shahmohammadi* 117  
*Gilberto N. Neto* 131  
*Hanaa Hachimi* 87  
*Harald Richter* 23  
*Harald Sporer* 183  
*Ignacio Silva-Lepe* 01  
*Inchan Paek* 141  
*Inchan Paek* 151  
*Ivan S. Mitzev* 173  
*Jainesh Jaintylal Hira* 219  
*Jean Michel Ilie* 297  
*Jian Tan* 267  
*Jinbae Suh* 151  
*Jonghun Jang* 141  
*Jonghun Jang* 151  
*Joohwan Chun* 141  
*Joohwan Chun* 151  
*Jörg Schäfer* 307  
*Jorge F Hernandez* 01  
*Karim Said Barsim* 77  
*Manuel Avalos* 01  
*Marcin Michalak* 257  
*Marcus Vinicius Lima Batista* 131  
*Maryam Abbasi* 201  
*Milson L. Lima* 131  
*Mohd Sapiyan* 323  
*Mora, A* 159  
*Nabil Hmina* 87  
*Nacer Tabib* 297  
*Nadson S. Timbó* 131  
*Nickolas H. Younan* 173  
*Pandu Rangan C* 53  
*Pedro Furtado* 201  
*Pedro Martins* 201  
*Rakeshh Mohan Bhatt* 45  
*Reda Mohamed HAMOU* 13  
*Reinhard German* 241  
*Rodríguez-Jiménez, J. M* 159  
*Sanaa Tayb* 87  
*Sanghyouk Choi* 141  
*Sasirekha Kambhampaty* 35  
*Sebastian Siegl* 241  
*Sergej Alekseev* 307  
*Shankar Kambhampaty* 35  
*Shenghua Wang* 267  
*Sofiane Labidi* 131  
*Taher Ali* 323  
*Thiago P. do Nascimento* 131  
*Thierry Mbah Mbelli* 219  
*Victor M Larios* 01  
*Yamshanov Artem* 63  
*Yankovskaya Anna* 63  
*Youssef Azdoud* 87  
*Yu Shoujian* 287  
*Zhou Yiyang* 287  
*Ziad Najem* 323