David C. Wyld
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

5th International Conference on Artificial Intelligence and Applications (AIAPP 2018), February 24~25, 2018, Dubai, UAE

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

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

ISSN: 2231 - 5403
ISBN: 978-1-921987-82-3
DOI : 10.5121/csit.2018.80401 - 10.5121/csit.2018.80410

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

Preface

The 5th International Conference on Artificial Intelligence and Applications (AIAPP 2018) was held in Dubai, UAE during February 24~25, 2018. The 4th International Conference on Cryptography and Information Security (CRIS 2018), The 5th International Conference on Computer Science and Information Technology (CoSIT 2018), The 5th International Conference on Signal and Image Processing (SIGL 2018) and The 4th International Conference on Software Engineering (SEC 2018) was collocated with The 5th International Conference on Artificial Intelligence and Applications (AIAPP 2018). 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 AIAPP-2018, CRIS-2018, CoSIT-2018, SIGL-2018, SEC-2018 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, AIAPP-2018, CRIS-2018, CoSIT-2018, SIGL-2018, SEC-2018 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 AIAPP-2018, CRIS-2018, CoSIT-2018, SIGL-2018, SEC-2018.

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.

David C. Wyld
Dhinaharan Nagamalai

Organization

General Chair

David C. Wyld	Southeastern Louisisna University, USA
Jan Zizka	Mendel University in Brno, Czech Republic

Program Committee Members

Abbas Jalilvand	Islamic Azad University, Iran
Abbas Khosravi	Deakin University, Australia
Abdelrahman Osman Elfaki	University of Tabuk, KSA
Abderrahmane Nitaj	University of Caen Normandy, France
Abdulghani Ali Ahmed	Universiti Malaysia Pahang (UMP), Malaysia
Agoujil Said	University of Moulay Ismail Meknes, Morocco
Ahmad Qawasmeh	The Hashemite University, Jordan
Ahmed Mohamed Refaat Azmy	Tanta University, Egypt
Akbar Majidi	Shanghai Jiao Tong University, China
Alessio Ishizaka	University of Portsmouth, United Kingdom.
Ali Asghar Rahmani Hosseinabadi	Islamic Azad University Amol Branch, Iran
Ali Selamat	Universiti Teknologi Malaysia, Malaysia
Amin Seyyedi	Islamic Azad University, Iran
Amizah Malip	University of Malaya, Kuala Lumpur
Anazida Zainal	Universiti Teknologi Malaysia, Malaysia
Àngela Nebot	Universitat Politécnica de Catalunya (UPC), Spain
Asad Abdi	Universiti Teknologi Malaysia (UTM), Malaysia
Azah Kamilah Muda	Universiti Teknikal Malaysia Melaka, Malaysia
Badir Hassan	Abdelmalek Essaadi University, Morocco
Bouchra Marzak	Hassan II University, Morocco
Cheng-Chi Lee	Fu Jen Catholic University, Taiwan
Claude Fachkha	University of Dubai, UAE
Edwin Lughofer	Johannes Kepler University Linz, Austria
Eng. Sattar B. Sadkhan	University of Babylon, Iraq
Erman Çakit	Aksaray University, Turkey
Fuxbuestc	Hangzhou Dianzi University, P.R.China.
Gazi Erkan Bostanc	Ankara University, Turkey
Grigorios N. Beligiannis	University of Patras, Greece
Habibollah Haron	Universiti Teknologi Malaysia, Malaysia
Haffaf H.	Département Informatique, Algérie
Haibo Yi	Shenzhen Polytechnic, China
Haitham Samy Elwahash	Kafrelshikh University, Egypt
Hamid Ali Abed AL-Asadi	Basra University, Iraq
Hamid Rastegari	Islamic Azad University, Iran
Hamido Fujita	Iwate Prefectural University, Japan
Hani Bani-Salameh	Hashemite University, Jordan
Hedieh Sajedi	University of Tehran, Iran

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Networks & Communications Community (NCC)

Information Technology Management Community (ITMC)

Organized By

Academy & Industry Research Collaboration Center (AIRCC)

TABLE OF CONTENTS

5th International Conference on Artificial Intelligence and Applications (AIAPP 2018)

Data Science Methodology for Cybersecurity Projects..... 01 - 14
Farhad Foroughi and Peter Luksch

GRC-MS: A Genetic Rule-Based Classifier Model for Analysis of Mass Spectra Data..... 15 - 36
Sara Al-Osimi and Ghada Badr

Understanding People Title Properties to Improve Information Extraction Process..... 37 - 46
Saleem Abuleil and Khalid Alsamara

5th International Conference on Computer Science and Information Technology (CoSIT 2018)

Strategy of The Remove and Easy TBT in GCC 6 Countries..... 47 - 52
Yong-Jae Kim

Distributed System Approach to Experiment Regional Competitiveness..... 103 - 109
Mhamed Itmi and Abdelkhalak El Hami

Reliability of Mechanical System of Systems..... 111 - 120
El Hami Abdelkhalakl and ITMI Mhamed

4th International Conference on Cryptography and Information Security (CRIS 2018)

Performance Analysis of Symmetric Key Ciphers in Linear and Grid Based Sensor Networks..... 53 - 68
Kaushal Shah and Devesh C. Jinwala

Perturbed Anonymization : Two Level Smart Privacy for LBS Mobile Users.. 69 - 79
Ruchika Gupta, Udai Pratap Rao and Manish Kumar

5th International Conference on Signal and Image Processing (SIGL 2018)

A Proposed HSV-Based Pseudo-Coloring Scheme for Enhancing Medical Images..... 81 - 92
Noura A. Semary

4th International Conference on Software Engineering (SEC 2018)

An Ontology Based Data Warehouse for the Grain Trade Domain..... 93 - 101
Mhamed Itmi and Boulares Ouchenne

DATA SCIENCE METHODOLOGY FOR CYBERSECURITY PROJECTS

Farhad Foroughi and Peter Luksch

Institute of Computer Science University of Rostock, Rostock, Germany

ABSTRACT

Cybersecurity solutions are traditionally static and signature-based. The traditional solutions along with the use of analytic models, machine learning and big data could be improved by automatically trigger mitigation or provide relevant awareness to control or limit consequences of threats. This kind of intelligent solutions is covered in the context of Data Science for Cybersecurity. Data Science provides a significant role in cybersecurity by utilising the power of data (and big data), high-performance computing and data mining (and machine learning) to protect users against cybercrimes. For this purpose, a successful data science project requires an effective methodology to cover all issues and provide adequate resources. In this paper, we are introducing popular data science methodologies and will compare them in accordance with cybersecurity challenges. A comparison discussion has also delivered to explain methodologies' strengths and weaknesses in case of cybersecurity projects.

KEYWORDS

Cybersecurity, Data Science Methodology, Data-Driven Decision-making, User Data Discovery, KDD Process, CRISP-DM, Foundational Methodology for Data Science, Team Data Science Process

1. INTRODUCTION

Cybersecurity solutions are traditionally static and signature-based which means they depend on pattern identification by detecting a match between a pre-captured attack or a malware with a new threat [1]. Hence, it needs regular update to deploy new signatures in the product database. For this reason, it is not possible to detect or prevent zero-day attacks.

The traditional solutions along with the use of analytic models, machine learning and big data could be improved by automatically trigger mitigation or provide relevant awareness to control or limit consequences of threats. Furthermore, traditional solutions are very binary with limited advantages compared to predictive models that could predict the possibility of attacks or risky actions based on data analysis techniques. In addition, access to a large amount of data makes it possible to solve challenging and complicated security problems. In accordance with big data and data mining, the more data creates more accurate and precise analysis [1].

This kind of intelligent solutions is covered in the context of Data Science for Cybersecurity. A general definition of data science is the information extraction and knowledge discovery from data by using a scientific approach. Data science could build innovative cybersecurity solutions by utilising new technological advantages of storage, computing and behavioural analytics [2]. In general, cluster storages which are deployed by distributed systems make it easier to collect and store huge amount of data (big data), and cloud computing also make it possible to utilise

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 01–14, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80401

complex and sophisticated machine learning techniques to create predictive and analytic models to identify and detect and respond threats. Behavioural analytics also could transform traditional signature-based detection techniques to the new behaviour-based predictive solutions.

According to data analysis potentials in cybersecurity, National Institute of Standards and Technology has developed a framework consists asset risk identification (and threat consequences), information protection, intruders detection, responding to intruders and business recovery [3].

As a result, Data Science provides significant role in cybersecurity by utilising the power of data (and big data), high-performance computing and data mining (and machine learning) to protect users against cybercrimes. For this purpose, a successful data science project requires an effective methodology to cover all issues and provide adequate resources.

In this paper, we are introducing popular data science methodologies and will compare them in accordance with cybersecurity challenges. Section 2 describes general Data Science overview along with its relation to the cybersecurity concept. Section 3 provides information about popular data science methodologies in details. Four different methodologies have been explained in this section. A comparison discussion is delivered in section 4 to explain methodologies' strengths and weaknesses against each other along with a summary table. In the end, we recommend a methodology that might cover all requirements to provide the most possible efficient data science cybersecurity project.

2. DATA SCIENCE

Data science could enhance and improve decision-making process by providing data-driven predictions. This requires principles, processes and techniques to understand a problem through an automated evaluation and data analysis. A successful data science has to employ data mining algorithms to solve business problems from a data perspective [4].

Data science is a set of basic concept which leads to the principled extraction of information and knowledge from data. It is very similar to the data mining tries to extract this information via technologies and applied and utilised for relationship management and behaviour analysis in order to recognize patterns, values and user interests [4]. CFJ Wu identified differences between traditional pure statistics and modern data science practices in 1997. He described these significant factors as Data Collection, Data Modelling and Analysis, and Problem Solving and Decision Support [5]. Data science is a recursive process which requires iterative performing.

Fayyad says Data Mining is a component of knowledge discovery in database process [6]. The knowledge discovery in database (KDD) was composed in 1989 to refer to the wide practice of obtaining and acquiring knowledge in data to stress the high-level application of certain data mining techniques. According to Fayyad et al. definition, KDD is the process of utilising data mining techniques to draw out knowledge based on the specification of measures and threshold, making use of a database together with any necessary pre-processing, sampling and transforming the database [7]. Therefore, a knowledge discovery process requires at least, Selection, Pre-processing, Transformation, Data mining, Interpretation and Evaluation steps. Knowledge discovery in database in cybersecurity domain interpreted into two major concepts. These concepts are User Data Discovery (UDD) with is a user profiling process and Data-Driven Decision-making which is a decision-making process based on data analysis [8].

2.1. Data-Driven Decision-making

Data-driven decision-making (DDD) is the term for the process and technique of taking decisions based on data analysis and information evaluation rather than strictly on intuition [4].

DDD is not a binary practice to provide all or nothing. It could be employed in cybersecurity domain with different levels of engagement. Provost et al. demonstrate two types of decisions: 1) decisions which are based on data discovery 2) decisions which are based on frequent decision-making processes particularly at considerable dimensions or massive scale. This kind of decision-making processes might gain from an even minor increase in reliability and precision based on information evaluation and data analysis [4].

Figure 1 describes data science and data-driven decision-making relation. Data Science overlaps data-driven decision-making because cybersecurity decisions and choices could increasingly be created instantly and automatically by computer systems [4].

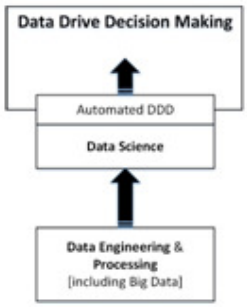


Figure (1): Data-Drive Decision-making through Data Science

Data processing and data engineering are essential to support data science tasks and very beneficial for data-driven decision-making, effective transaction processing and online pattern recognition. Big Data is simply a term for datasets which are very large for conventional data processing and require new methods and technologies. Therefore, big data technologies are in fact utilised for applying data processing to support data mining strategies and data-driven decision-making tasks [4]. Modern efficient cybersecurity solutions are depend on big data because more data creates more accurate and precise analysis [1].

Data analytic thinking is a crucial element of data science. Underlying the comprehensive collection of methods and strategies for mining information is a significantly smaller set of basic concepts comprising data science. Understanding the essential concepts and having a data analytic thinking structure and framework could help cybersecurity researchers to boost data-driven decision-making process.

2.2. User Data Discovery

User Data Discovery (UDD) is the process of producing profile of users from historical information and past details. This particular information might be personal data, academic documents, geographical details or other private activities.

The primary function of user profiling process is capturing user's information about interest domain. These information may be used to understand more about individual's knowledge and skills and to improve user satisfaction or help to make a proper decision. Typically, it evolves data mining techniques and machine learning strategies. UDD process is a type of

knowledge discovery in database or the new version, knowledge data discovery model and requires similar steps to be established [8].

User profiling is usually either knowledge-based or behaviour based. The knowledge-based strategy uses statistical models to categorise a user in the closest model based on dynamic attributes. Typically, Questionnaires and interviews could be utilised to acquire this particular user knowledge [9].

Behaviour-based strategy employs the user's behaviours and actions as a model to observe beneficial patterns by applying machine learning techniques. These behaviours could be extracted through monitoring and logging tasks [9].

Recognizing user behaviour in real time is an important element of providing appropriate information and help to take suitable action or decision in cybersecurity projects. Typically it is a human task that experts would provide in the information security domain, but it is possible to employ user modelling to make this process automatic by using an application or intelligent agent [10].

UDD could obtain appropriate, adequate and accurate information about a user's interests and characteristics and demonstrate them with minimal user intervention [11] to offer appropriate awareness with relevant mitigation recommendation based on the security situation.

An intelligent cybersecurity solution should take into account the various attributes and features of a user and a security situation to create a customized solution based on the notion and concept of user profile [12].

3. DATA SCIENCE METHODOLOGY

Several theoretical and empirical researchers have considered basic concepts and principles of knowledge extraction from data. These basic methods and fundamental principles are concluded from numerous data analytic studies [4].

Extracting beneficial knowledge from data should be dealt with systematic processes and procedures through well-defined steps.

Data science needs attentive consideration and result evaluation in the context it is used because the extracted knowledge is significant to assist the decision-making process in a particular application [4].

"Breaking the business problem up into components corresponding to estimating probabilities and computing or estimating values, along with a structure for recombining the components, is broadly useful." [4].

The correlation finding is one of the data science concepts which should be considered in relation to the cybersecurity. It typically provides details on data items that supply information about other data items, particularly, recognized quantities which reduce the uncertainty of unknown quantities [4].

Entities which are identical with regard to known features or attributes, oftentimes are identical with regard to unknown features or attributes. Computing similarity (pattern recognition) is among the primary resources of data science [4]. It is also significant to pay quite close attention to the existence of confounding elements, most likely unseen ones.

A methodology is a general approach that guides the techniques and activities within a specific domain. The methodology does not rely on certain technologies or tools. Instead, a methodology delivers a framework to acquire results by using a wide range of methods, processes and heuristics [13].

Predictive model creation, pattern recognition and underlying discovery problems through data analysis are usually a standard practice. Data science provides plenty of evolving data analysis technologies to constructing these models. Emerging analytics methods and action automation provide strong machine learning models to solve sophisticated analytic problems such as DDDs. To create an appropriate data analytic model it is required to use a data science methodology that could provide and supply a guiding strategy regardless of technology, data volumes or approaches.

There are several methodologies available for data mining and data science problems such as CRISP-DM and SAS SEMMA and KDD process but Gregory Piatetsky confirms CRISP-DM remains the top methodology for data mining projects with 42% in 2014. The KDD Process has used by 7.5% [14].

Rollins demonstrates a foundational methodology which is similar to CRISP-DM, SEMMA and KDD Process but also emphasizes a number of new methods in data science including big data usage, the incorporation of text analytics into predictive modelling and process automation [13]. Microsoft also introduces Team Data Science Process (TDSP) which recommends a lifecycle for data science projects [15].

Before applying any of these methodologies to cybersecurity projects, it might be helpful to review and compare their essential features. For this reason, this paper provides a comparison between KDD Process, CRISP-DM, TDSP and the foundational methodology for data science (FMDS). FMDS and CRISP-DM have been chosen, because they are considered to be the most popular but SAS SEMMA is not in this review because there is a big decline in applying it (from 13% in 2007 to 8.5% in 2014) [14]. The KDD Process has also included because it provides initial and basic requirements of knowledge discovery. TDSP has been chosen because it is customized for machine learning or artificial intelligence projects which are considerably linked to cybersecurity applications.

3.1. KDD Process

The KDD process offered by Fayyad et al. in 1996 [7]. It is the method of using data mining techniques to extract knowledge based on particular measures and thresholds in a database by employing any necessary pre-processing, sampling or data transformation actions [7]. Furthermore, the application domain perception is needed to be considered during the KDD process development, improvement and enhancement. Figure 2 illustrates the KDD process.

The KDD process has 5 steps as the following [7].

1. Selection: It means generating or producing a target data set or concentrating on a subset of variables or data samples in a database.

2. Pre-processing: This phase tries to obtain consistent data by cleaning or pre-processing selected data.

3. Transformation: Reducing feature dimensionality by using data transformation methods is performing in this phase.

4. Data Mining: Trying to recognize patterns of attention or behaviours by using data mining techniques in a specific form should perform in this step. (Typically, prediction)

5. Interpretation/Evaluation: Mined pattern should be assessed and interpreted in the final phase.

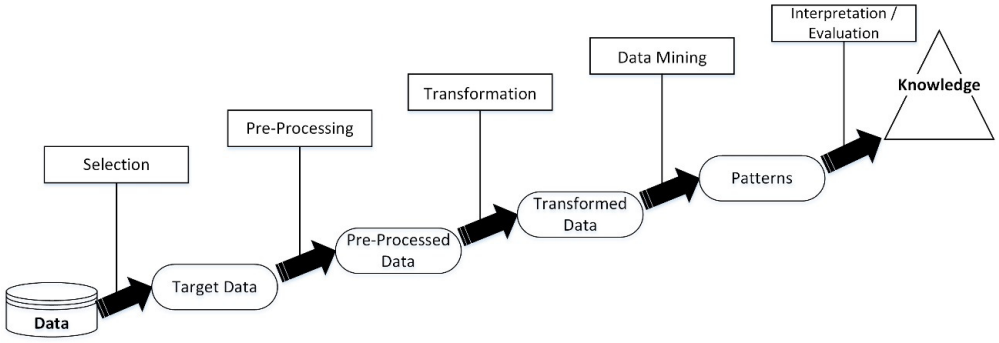


Figure (2): The KDD Process

3.2. CRISP-DM

In 1996, SPSS and Teradata developed Cross Industry Standard Process for Data Mining (CRISP-DM) in an effort initially composed with NCR, Daimler and OHRA. It is a cycle of six high-level phases which describe the analytic process [16, 17].

CRISP-DM is still a beneficial tool but details and specifics needed to be updated for cybersecurity projects such as those including Big Data. The original site is not active anymore but IBM SPSS modeller still supports it [14].

Figure 3 demonstrates the CRISP-DM six stages, but their sequence is not strict. CRISP-DM is very well documented and there are many case studies which have used it [16].

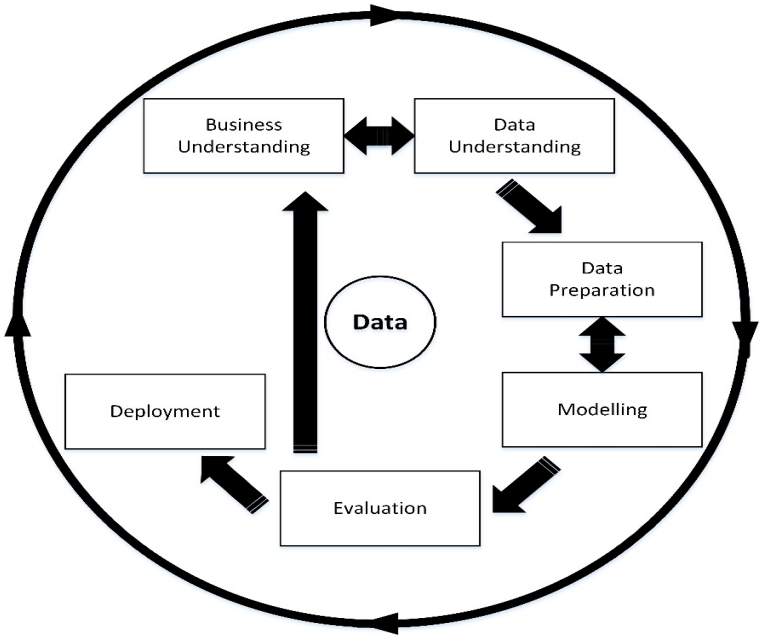


Figure (3): The CRISP-DM life cycle

CRISP-DM's structured, well defined and extremely documented process is independent of data mining tools and this important factor is crucial to make the project successful [16].

The CRISP-DM is so logical and seems like common sense. There are many methodologies and advanced analytic platforms which are actually based on CRISP-DM steps because the use of a commonly practised methodology gains quality and efficiency. Vorhies says CRISP-DM provides strong guidance for even the most advanced of today's data science projects [17].
The six phases are the following [16]:

1. Business understanding: It is designed to focus on understanding the problem or project goals and requirements from a business perspective (here is a cybersecurity application) and then transforming this perception into data mining problem description and preliminary approach.

2. Data Understanding: This phase begins with an initial data collection and then with tasks in order to acquaint with data, to recognise data quality problems, to find out primarily insights into the details or even to identify interesting subsets to develop hypotheses for hidden information. It typically creates a set of raw data.

3. Data Preparation: This phase covers all tasks and activities to build the final required dataset from the first raw data.

4. Modelling: Through this stage, modelling techniques and strategies are selected and applied, and their specific parameters and prerequisites should be identified and calibrated regarding the type of data to optimal values.

5. Evaluation: At this point, the obtained model or models which seem to provide high quality based on loss function completely evaluated and the actions executed to ensure they generalise against hidden data and to be certain it correctly archives the key business goals. The final result is the selection of sufficient model or models.

6. Deployment: This stage means deploying a code representation of the final model or models in order to evaluate or even categorise new data as it arises to generate a mechanism for the use of new data in the formula of the first problem. Even if the goal of the model is to provide knowledge or to understand the data, the knowledge acquired have to be organised, structured and presented in a means which could be used. This includes all the data preparation and required steps which are needed to treat raw data to achieve the final result in the same way as developed during model construction.

3.3. Foundational Methodology for Data Science

This methodology has some similarities and consists many features of KDD Process and CRISP-DM, but in addition, it provides a number of new practices such as use of extremely large data volumes, text and image analytics, deep learning, artificial intelligence and language processing [13]. The FMDS's ten steps illustrate an iterative nature of the problem-solving process for utilising data to discover security insights. Figure 4 demonstrates FMDS process.

Foundational Methodology for Data Science consists the following steps [18]:

1. Business understanding: This very first phase provides a basic foundation for a profitable and effective resolution of a business problem (a cybersecurity challenge in here) regardless of its size and complexity

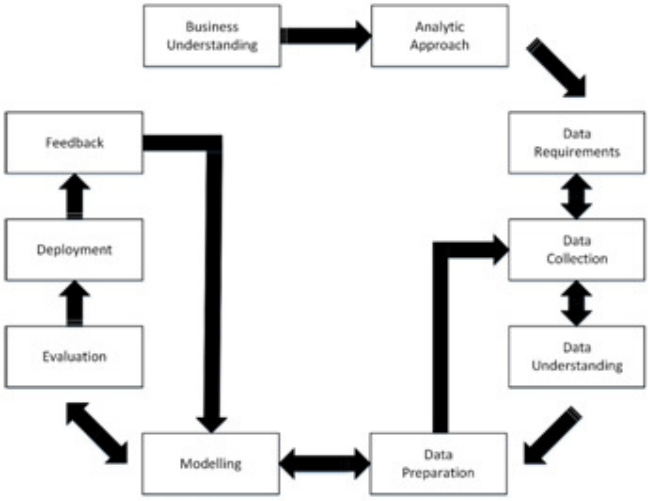


Figure (4): Foundational Methodology for Data Science

2. Analytic approach: As soon as the problem clearly stated, it is required to determine an analytic approach by identifying suitable machine learning technique to solve it and obtain the desired result.

3. Data requirements: The analytic approach that has been chosen in the second phase, defines needed data requirements including specific data content, formats and representations which are instructed by cybersecurity knowledge.

4. Data collection: In this primary data gathering phase, it is required to identify and collect available data resources (structured, unstructured and semi-structured) which are related and applicable to the problem domain. It is important to cover any data collection gap by revising data requirements and gathering brand new and/or additional data.

It is also significant to use high-performance platforms and in-database analytic functionality to work with huge datasets for sampling and sub-setting to obtain all available data.

5. Data understanding: Descriptive statistics and visualisation methods are useful in this phase to understand data content, evaluate data quality and explore data insights. This could be required to revise the earlier phase to close data collection gaps.

6. Data preparation: This phase comprises all tasks to construct dataset which will be utilised in the modelling phase. These tasks include data cleaning, data merging from several sources, dealing with missing data, data transformation into more useful variables, duplication elimination, and finding invalid values. In addition, feature engineering and text analytics are possible to be utilised to provide new structured variables, defining and enriching the predictors and boosting or improving the model's reliability, accuracy and precision. A collaboration of cybersecurity knowledge and existing structured variables could be very useful for feature engineering. This phase is probably the most time-consuming stage, but high-performance and parallel computing systems could reduce the time required and prepare data quickly from huge datasets.

7. Modelling: This phase concentrate on predictive or descriptive model development based on earlier described analytic approach and by using the first version of the prepared dataset as a training set (historical data). The modelling process is extremely iterative as it provides

intermediate insights and reputable refinement of data preparation and model specification. It is significantly helpful to try several algorithms with specific parameters to find the ideal model.

8. Evaluation: Before deployment, it is crucial to evaluate the quality and efficacy of the developed model to realise whether it completely and appropriately addresses the cybersecurity problem. This evaluation involves computing of several different diagnostic measures and other outputs including tables and graphs by using a testing set. This testing set is actually independent of the training set but follows the identical probability distribution and has known results.

9. Deployment: Once the developed model approved and accredited in the evaluation phase that covers the cybersecurity challenge appropriately, it should be deployed into the production environment or perhaps in a comparable test environment. Typically, it will be used in a limited and specific way until all performance variables entirely evaluated. Deployment could be as basic as producing a simple report with proper suggestions or perhaps, embedding the model in an elaborated or sophisticated workflow and scoring process handled by a customised application.

10. Feedback: This final phase, collects outcomes from the implemented edition of the analytic model to analyse and feedback its functionality, performance and efficiency in accordance with the deployment environment.

3.4. Team Data Science Process

The Team Data Science Process (TDSP) is a data science methodology to provide efficient predictive analytics. TDSP These solutions are including artificial intelligence and machine learning. It boots data science project agility, team working and learning by employing best practices and successful structures from Microsoft [15]. TDSP supports both exploratory and ad-hoc projects. Figure 5 illustrates TDSP five stages.

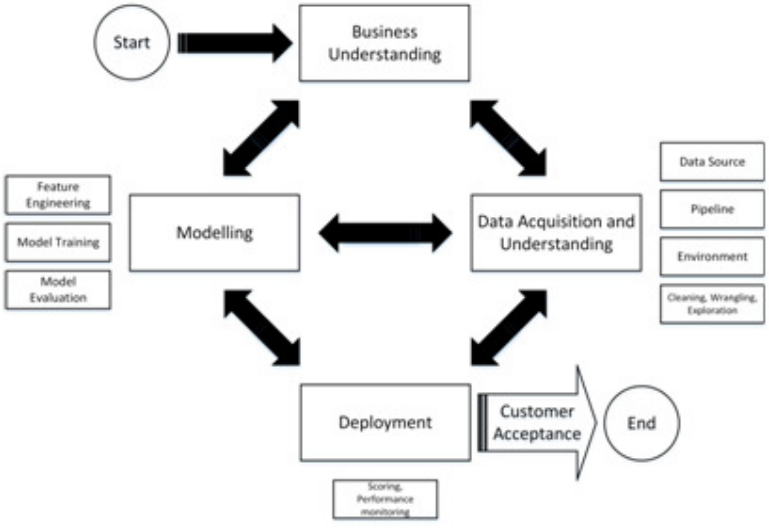


Figure (5): Team Data Science Process Lifecycle

TDSP supports development of projects which have already employed CRISP-DM and KDD process. It is very customizable based on project's size and dimensions [19].

The TDSP lifecycle consists integrative phases as the following [19].

1. Business Understanding: Initially, a question which describes the problem objectives should be defined clearly and explicitly. Relevant predictive model and required data source/s also have to be identified in this step.

2. Data Acquisition and Understanding: Data collection starts in this phase by transferring data into the target location to be utilised by analytic operations. The raw data is needed to be cleaned and incomplete to incorrect values should be identified. Data summarization and visualization might help to find required cleaning procedures. Data visualization also could help to measure if data features and collected amount of data are adequate over the time period. At the end of this stage, it might be necessary to go back to the first step for more data collection.

3. Modelling: Feature engineering and model training are two elements of this phase. Feature engineering provides attributes and data features which are required for machine learning algorithm. Algorithm selection, model creation and predictive model evaluation are also sub-components of this step. Collected data should be divided into training and testing datasets to train and evaluate machine learning model. It is important to employ different algorithms and parameters to find the best suitable solution to support the problem.

4. Deployment: Predictive model and data pipeline are needed to be produced in this step. It could be a real-time or a batch analysis model depends on the required application. The final data product should be accredited by the customer.

5. Customer Acceptance: The final phase is customer acceptance which should be performed by confirming data pipeline, predictive model and product deployment.

4. DISCUSSION

All data science methodologies where discussed consist four common iterative stages including problem definition/ formulation, data gathering, data modelling and data product development except the KDD process.

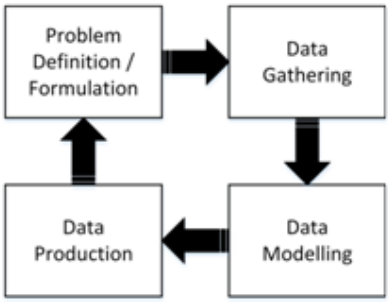


Figure (6): General Data Science Methodology

According to the above explanations, table 1 demonstrates a summary of the presented correspondences.

Comparing the KDD process with the CRISP-DM presents that KDD process does not cover the business understanding and also deployment phases in CRISP-DM methodology.

As it is mentioned above, the business understanding phase provides a comprehension perception of the application domain and pertinent prior knowledge and also objectives of required solution. Furthermore, the deployment phase incorporates knowledge or modelling code into a system or

application to build a data product. These two significant phases have been missed in KDD process.

Table (1): Summary of Data Science methodologies and their phases

KDD Process	CRISP-DM	FMDS	TDSP
-	Business Understanding	Business Understanding	Business Understanding
	-	Analytic Approach	
	-	Data Requirements	
Selection	Data Understanding	Data Collection	Data Acquisition and Understanding
Pre Processing		Data Understanding	
Transformation	Data Preparation	Data Preparation	
Data Mining	Modelling	Modelling	Modelling
Interpretation/ Evaluation	Evaluation	Evaluation	
-	Deployment	Deployment	Deployment
-	-	Feedback	Customer Acceptance

Comparing the CRISP-DM with the FMDS illustrates that CRISP-DM has not the analytic approach and the feedback phases. The analytic approach is required to recognize appropriate statistical or machine learning techniques before entering to the data gathering steps. This phase could be very useful to identify the suitable data collection strategy and data resources but it is missed in the CRISP-DM methodology. The feedback phase also has been missed that is very beneficial to optimize the system to achieve high-performance functionality and efficient result.

A comparison between FMDS and TDSP presents that they are very similar but FMDS has more details steps in general. Feedback also is a part of FMDS cycle which could create new requirements to improve the data product in an iterative process but customer acceptance is a stage out of the circle in the TDSP. FMDS detailed stages could be more useful for a wide range of projects but TDSP uses a specific set of Microsoft tools and infrastructure to deliver intelligent applications by deploying machine learning or AI models.

Concerning the remaining phases present the following:

- The Data understanding phase is in both CRSIP-DM and FMDS can be recognized as the collaboration of Selection (Collection) and Pre-Processing phases in the KDD process but Data Acquisition and Understanding in TDSP also covers Transformation stage in the KDD process. However, data requirements which are related to the analytic approach phase and provides required data content are missed in both KDD process and CRISP-DM. Selecting an analytic approach is integrated into the business understanding phase of TDSP but cybersecurity projects might gain more benefits in details from this task when it is an independent step in the FMDS particularly when data resources are separated and requires a different method or level of access.

- Business understanding and problem formulation is an initial phase that makes the data understanding phase more efficient by recommending data formats and representations but data requirement analysis is missed in the CRISP-DM. It might be very crucial in cybersecurity projects particularly when data resources are unstructured or semi-structured.

- The data preparation phase has the similar function as the transformation phase in KDD process and it is included in the Data Acquisition and Understanding phase of TDSP by using some tools.

- The modelling phase might be recognised with data mining phase which is very limited in the KDD process. Modelling phase in TDSP also has the evaluation task included by using some tools such as scoring and performance monitoring, but it is an independent phase in other methodologies.

- The evaluation phase is also included in all three methodologies.

In spite of CRISP-DM strengths, there is more emphasis that should be considered for modern cybersecurity projects.

- It is required that methodology could also handle data blending from several sources. This should be deployed through a completely repeatable process. FMDS and TDSP provide this feature in the Data Preparation and Data Acquisition and Understanding phases. TDSP provides Microsoft tools to support On-Premises, Cloud, Databases and Files but FDMS is independent of any platform or a specific tool. This might makes cybersecurity projects more reliable and efficient.

- Choosing the most appropriate degree of reliability and accuracy for the problem is very crucial to make sure there is no need to spend excessive time on data preparation or modelling to enhance accuracy when it could not be utilised. This feature also included in the FMDS methodology.

- The entire analytic algorithm should be tested and evaluated to make sure there are working in all situations and not just for sample modeller. The evaluation, deployment and feedback cycle in the FMDS could provide this requirement as well as Model training task of Modelling phase in the TDSP. Feedback phase in FMDS might create new data science questions to optimize the cybersecurity product or make new functionalities for it.

- It is significant to consider quality during model simplification by ensuring that decision elements such as missing data reduction, synthetic features generation and unseen data holdout are properly managed. The evaluation, deployment and feedback cycle in the FMDS could bring this need better than simple quality insurance in the evaluation phase of CRISP-DM.

- Data science lifecycle is very well defined in the FMDS and connections are clearly determined between every stage but TDSP's stages are all linked together (except customer acceptance) and it is possible to move into any stages from anyone else.

- TDSP lifecycle is designed for predictive analytic projects by using machine learning or artificial intelligence models. FMDS is more general and independent of any platform, tool, model or algorithm. Both are functional for exploratory or ad-hoc cybersecurity analytics projects by some customization.

5. CONCLUSION

In conclusion, a cybersecurity data science project requires four general steps. The first step has to be a problem definition by formulating a security challenge. In accordance with problem definition and appropriate formula, it is necessary to gather required information in the second step. The collected information should be employed in the third step and in an analysis process to provide adequate data which is expected to predict or provide a resolution for the defined

problem. The final step is a production step which deploys relevant modules and a system to run the whole process automatically and regularly when it is needed.

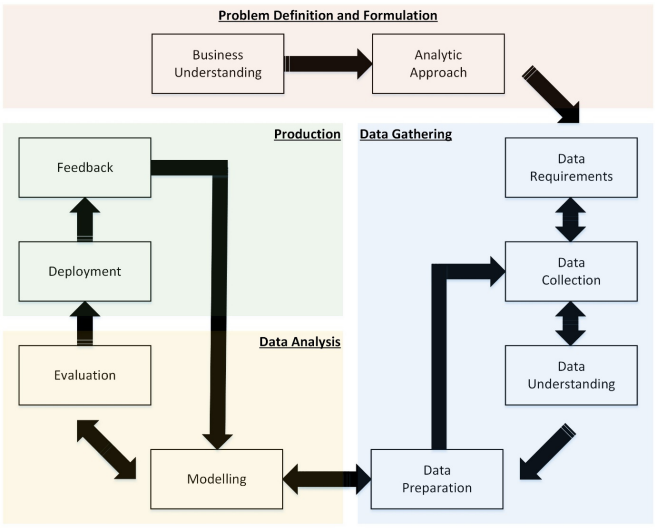


Figure (7): FMDS for Cybersecurity projects

Regarding the general process and in accordance to Table 1, the FMDS covers all beneficial attributes of CRISP-DM methodology but fills data gathering gaps and also provides extra steps to optimize and enhance the model and results by mathematical prescriptive analytics and using high-performance computing. It is also platform and tool independent but TDSP is not. Because it is designed in details with clearly divided steps, it could be fully customized to fit in any cybersecurity project.

REFERENCES

[1] R. Mastrogiacomo, "The conflict between data science and cybersecurity," Available: https://www.information-management.com/opinion/the-conflict-between-data-science-and-cybersecurity

[2] D. L. Pegna, "Creating cybersecurity that thinks," Available: https://www.computerworld.com/article/2881551/creating-cyber-security-that-thinks.html

[3] "Framework for Improving Critical Infrastructure Cybersecurity," National Institute of Standards and Technology, 2014.

[4] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," Big Data, vol. 1, no. 1, pp. 51-59, 2013.

[5] C. Wu, "Statistics= Data Science?(1997)," ed, 1997.

[6] U. M. Fayyd, G. P. Shapiro, and P. Smyth, "From data mining to knowledge discovery: an overview," 1996.

[7] U. M. Fayyad, "Data mining and knowledge discovery: Making sense out of data," IEEE Expert: Intelligent Systems and Their Applications, vol. 11, no. 5, pp. 20-25, 1996.

[8] S. Kanoje, S. Girase, and D. Mukhopadhyay, "User profiling trends, techniques and applications," arXiv preprint arXiv:1503.07474, 2015.

[9] S. E. Middleton, N. R. Shadbolt, and D. C. De Roure, "Ontological user profiling in recommender systems," ACM Transactions on Information Systems (TOIS), vol. 22, no. 1, pp. 54-88, 2004.

[10] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, "Creating evolving user behavior profiles automatically," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 854-867, 2012.

[11] B. S. Atote, T. S. Saini, M. Bedekar, and S. Zahoor, "Inferring emotional state of a user by user profiling," in Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on, 2016, pp. 530-535: IEEE.

[12] S. Ouaftouh, A. Zellou, and A. Idri, "User profile model: a user dimension based classification," in Intelligent Systems: Theories and Applications (SITA), 2015 10th International Conference on, 2015, pp. 1-5: IEEE.

[13] J. B. Rollins, "Foundational Methodology for Data Science," Available: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=IMW14824USEN

[14] G. Piatetsky, "CRISP-DM, still the top methodology for analytics, data mining, or data science projects," KDD News, 2014.

[15] W. A. R. Roald Bradley Severtson, "What is the Team Data Science Process?," Available: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

[16] A. I. R. L. Azevedo and M. F. Santos, "KDD, SEMMA and CRISP-DM: a parallel overview," IADS-DM, 2008.

[17] W. Vorhies, "CRISP-DM – a Standard Methodology to Ensure a Good Outcome," vol. 2017, ed: Data Science Central, 2016.

[18] J. B. Rollins, "Why we need a methodology for data science," vol. 2017, ed: IBM big data and analytics hub, 2015.

[19] G. Kumar, "Team Data Science Process Lifecycle," Available: https://github.com/MSFTImagine/Microsoft-DataScience-Process/blob/master/Docs/team-data-science-process-lifecycle-detail.md

GRC-MS: A GENETIC RULE-BASED CLASSIFIER MODEL FOR ANALYSIS OF MASS SPECTRA DATA

Sara Al-Osimi[1] and Ghada Badr[2]

[1]Department of Computer Science, Shaqra University, Riyadh, KSA
[2]IRI - The City of Scientific Research and Technological Applications, Alex, Egypt

ABSTRACT

Many studies uses different data mining techniques to analyze mass spectrometry data and extract useful knowledge about biomarkers. These Biomarkers allow the medical experts to determine whether an individual has a disease or not. Some of these studies have proposed models that have obtained high accuracy. However, the black-box nature and complexity of the proposed models have posed significant issues. Thus, to address this problem and build an accurate model, we use a genetic algorithm for feature selection along with a rule-based classifier, namely Genetic Rule-Based Classifier algorithm for Mass Spectra data (GRC-MS). According to the literature, rule-based classifiers provide understandable rules, but not accurate. In addition, genetic algorithms have achieved excellent results when used with different classifiers for feature selection. Experiments are conducted on real dataset and the proposed classifier GRC-MS achieves 99.7% accuracy. In addition, the generated rules are more understandable than those of other classifier models.

KEYWORDS

Mass spectrometry, data mining, biomarkers, rule-based classifier, genetic algorithm.

1. INTRODUCTION

Mass spectrometry (MS) is an efficient technique that has been widely used in many disciplines, such as science, engineering, and biology. Recently, MS has been used in the bioinformatics field to identify the amounts of chemical and biological materials in human tissue or serum to use later as biomarkers. These biomarkers can be used as measures for clinical assessments to monitor and predict individuals' health conditions in order to plan suitable therapeutic interventions [1]. However, because the data generated using the MS technique is so huge and extensive, it is difficult to extract any useful knowledge or biomarkers; Many studies have been done to develop data mining analysis tools (i.e., classification, clustering, correlation analysis, etc.) for the interpretation and extraction of accurate knowledge from MS data. However, the results of most of these studies have not been satisfactory. Even when the studies do achieve good results, experts may struggle to understand them. According to the literature [2], rule-based classifiers yield acceptable results when they are applied to the analysis of discrete data. In addition, these

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 15–36, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80402

classifiers have the unique ability to provide very meaningful outcomes. However, unfortunately, rules-based classifiers do not achieve the quality required for analysis of MS data.

In this paper, we propose an efficient and meaningful approach that uses Genetic Algorithms (GAs), namely GRC-MS, to select features and then build a rule-based classification model with the objective of classifying and understanding MS data. We also test our proposed approach on a real dataset of ovarian cancer patients in order to measure the accuracy of the proposed approach. The proposed approach is intended to be a general framework that can be used for the analysis of any MS data or related continuous data. To the best of our knowledge, the combination of rule-based classifiers with GAs as the feature selection technique has not yet been applied to MS data.

This paper is organized as follows: Section 2 provides a background about the MS techniques: preprocessing, some feature selection, and classifiers that are used for MS data. Section 3refers to some of the studies that use GA technique as a feature selection approach for MS data. In addition, it summarizes some of the studies that use rule base techniques as classifiers for MS data. Section 4 explains the steps of our proposed approach, GRC-MS. The experimental setup and results on a real dataset are presented in Section 5.Section 6, discuss the results. Finally, Section 7 concludes the paper and discusses future work.

2. BACKGROUND

Mass spectrometry (MS) is a recently developed technique that is used to identify, analyze, and determine the elements, molecules, and atomic structures of any given sample [3]. MS quickly and accurately determines the relative numbers of molecules present in complex biological or chemical samples by transforming these samples into ions, detecting their mass-to-charge ratios (m/z), and then measuring the intensity of each ion type [4]. This technique is used primarily to study the effects of ionizing energy on sample molecules [3]. It has several beneficial characteristics, such as speed and sensitivity. Moreover, because MS has a variety of possible applications, it is preferable to other analytical methods and, as a result, has progressed rapidly over the last decade. Today, MS is used in a number of applications, such as biochemical problems, pollution control, atomic physics, food control, forensic science, reaction kinetics, geochronology, inorganic chemical analysis, process monitoring, and so on [4].

2.1 Proteomics

Proteomics, a term that is first coined by Australian scientist Marc Wilkins in 1994, is an emerging area in bioinformatics [7]. It provides information about proteins and their interactions in the human body. The major aim of most proteomic studies is the detection of proteins of interest, which are known as biomarkers. The term “biomarkers” refers to protein molecules that facilitate the detection of a particular cell type and that identify cell characteristics, such as cells’ ability to perform their functions. [8]. The discovery of biomarkers in MS data is useful for the early diagnosis of diseases. Most researchers hope to discover novel and powerful diagnostic proteomic tools to detect these biomarkers [8]. Recently, several techniques have been developed for analyzing bodily fluids, such as human serum, human urine, and, in some studies, tumor tissue, to achieve protein profiling. Commonly, the analysis of body fluids is accomplished using MS techniques [9]. Two major techniques are intended for proteomic analysis: MALDI and SELDI. MALDI-TOF MS is a new and widely used technique for discovering biomolecules, such as proteins with molecular masses between 400 and 350000 Da, in samples [6].

2.2 Mass Spectrometry

MS experiments are generally conducted in three main stages: the data generation stage, the data preprocessing stage, and the data analysis stage [4] [5]. In the first stage, MS techniques generate data that are represented as a huge sequences of pairs, called matrix, spectrum, or MS data [4]. This spectrum contains mass-to-charge ratio values and intensity values [6]. The mass-to-charge ratio values (which are represented on the x-axis) depend on the molecular mass detected in the sample, and the intensity values (which are represented on the y-axis) depend on the quantity of molecules detected in the sample (Figure 1) [6]. Depending on the resolution of the MS technique, a spectrum can contain hundreds or thousands of pair values [7]. Data preprocessing involves cleaning the data and improving their quality. On the other hand, during data analysis, data mining or pattern extraction techniques are applied to extract knowledge.

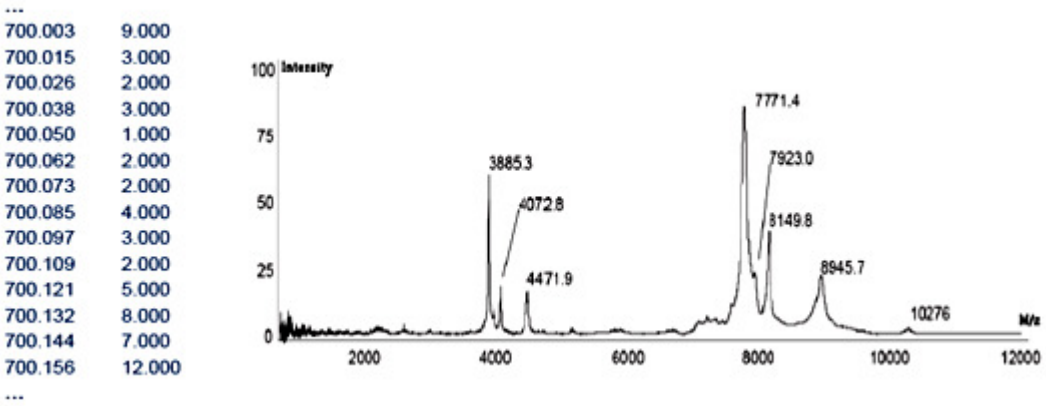


Figure 1. Output signals from a mass spectrometer consisting of m/z and intensity values [9].

2.3 Data Mining in Mass Spectrometry

Data mining is a well-known approach that is used in science and business to extract useful information from large and complex datasets [7][2]. The steps involved in data mining include (Figure 2) data preparation, feature selection, model development (or pattern recognition), and model assessment. The following section focuses on the basic algorithms used in data mining for application to mass proteomic data. However, as previously mentioned, MS data are high in dimensionality, and they cannot be analyzed through the direct use of data mining techniques. Preprocessing the MS data is a crucial step in improving the data quality—and, thus, improving the quality of the classifier algorithms [6].

- **Preprocessing MS Data**

MS data or spectra are commonly influenced by errors or noise that occur during the sample preparation or the insertion into the device or by noises generated by the device itself [4]. Using the raw MS data directly for the analysis process is not effective because contaminants like noise, m/z measurement errors, and matrix size affect the results [6] [7]. In addition, because of the dimensional complexity of the spectra, efficient results cannot be obtained through the direct application of data mining algorithms or pattern extraction techniques. Therefore, cleaning the MS data is critical. To achieve clean data, different preprocessing techniques are applied to the MS data before the application of any data mining technique such as reducing noise and

smoothing data, normalization, data reduction by binning, peak extraction, and peak alignment. These techniques can be used alone or in combination [10].

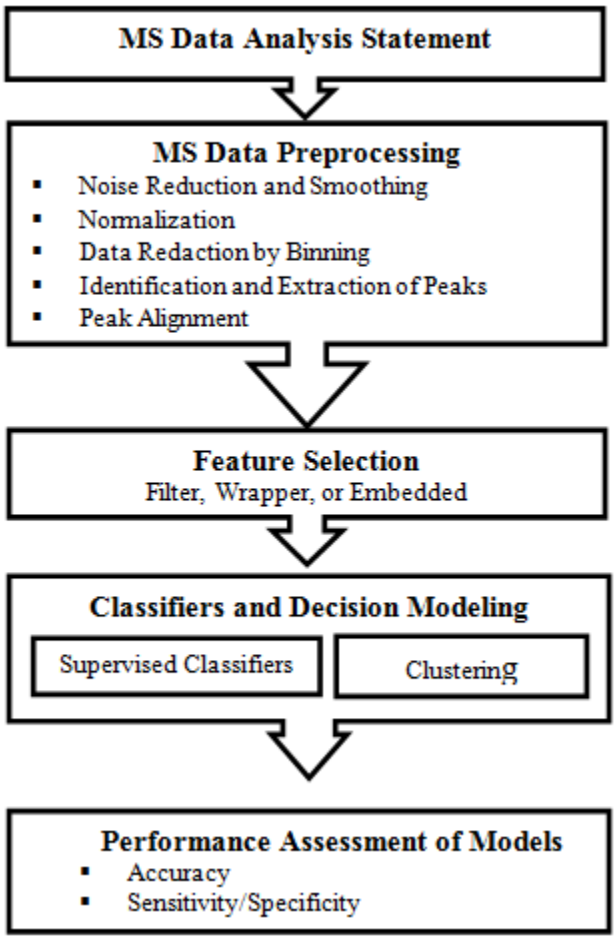


Figure 2. Typical flowchart of the critical steps in data mining and examples of the techniques available for MS data analysis.

- **Feature Selection Techniques**

The MS technique produces high-dimensional data. Compared to the number of samples, a greater number of peaks needs to be analyzed (high features-to-sample ratio datasets) [11]. Most of the problems in analyzing data stem from the size and the complexity of the datasets that are represented in tables of rows and columns. Rows represent records or cases, and columns represent data dimensions, features, and attributes [7]. In the analysis of MS data, the extraction uses the intensity of every peak in the spectrum as a feature. The number of features (or peaks) is usually large (e.g., 17,000 peaks), while the number of samples is usually small (e.g., 140 patients) [12]. However, features often contain noise with a very little or no informational value. Thus, it is necessary to select features from a large set of those likely to be useful in predicting the outputs of interest. To solve this problem, after the data is pre-processed, a feature selection phase step is performed. This step aims to detect the main parts of the spectrum that might provide a

better understanding of the data's important features, which could be used in the analysis phase [11].

Feature selection techniques can be divided into three categories [12]: filter, wrapper, and embedded. The filter technique analyzes each feature independently and eliminates features one at a time based on how they correlate with the target. Feature subsets are selected based on evaluation criterion, such as information gains. This is a simple and quick process that is sometimes referred to as independent feature selection. Moreover, filter selection methods are relatively computationally efficient [11] [12]. Examples of independent feature selection techniques used with MS or high-dimensional data include: statistical tests (i.e., t-tests [13] [14], Wilcoxon tests [17], $\chi 2$ tests [18]), information gains [19], and so on. The wrapper techniques simultaneously analyze features in groups or subsets and build the analysis model [11] [12]. Classifiers are used to assess (several) features or feature subsets. Although the process is computationally busy and potentially very time-consuming, since this technique typically requires an evaluation of every scheme at every iteration, it discovers critical information that is typically lost in independent features analysis [16] [20]. Examples of wrapper feature selection techniques that are used with MS or high-dimensional data include: genetic algorithms [21] [22], sequential searches [23], and estimations of distribution algorithms [24]. In embedded techniques, the search for a best set of features is made into the classifier construction. They learn which set of features can best contribute to the accuracy during the creation of the model [12]. These techniques make no distinction between learning and feature selection. Embedded techniques have the advantage of including the interactions with the classification model, while simultaneously being far less computationally intensive than wrapper methods [12]. Examples of embedded feature selection techniques that can be used with MS or high-dimensional data include random forest techniques [25] and weight vector support vector machine techniques [26].

- **Classifiers and Decision Models for MS Data**

For MS Data, usually classification or supervised learning uses to predict or classify new cases. Where, previous knowledge about classes can be used to classify new cases. The previous knowledge is built using a training dataset, which includes input values and their output classes. In the training stage, the training dataset is used to define how the features are to be selected and combined to distinguish among the different classes. In the testing stage, the weighted features are applied to classify a new test dataset. The test dataset's class is not known, and the dataset has never before been seen by the model. If the model classifies new cases correctly, it is a good model. A wide range of algorithms, such as decision tree algorithms, SVMs, ANNs, and so on, have been developed for classification. In this subsection, we well indicate to some of well-known classifiers that used for MS data.

i. iDecision Tree (DT) Classifier

Decision tree (DT) a hierarchical tree structure model that is described as a set of rules, presented in a visual form that is very easy to understand. The DT was used by Vlahou et al. [28] to analyze the MS data and it is achieved 80% accuracy in discriminating between ovarian cancer patients and healthy controls. In addition, Su et al. [29] used DT to analyze the MS data and they obtained 85.3% accurate.

ii. Artificial Neural Networks (ANNs) Classifier

Artificial neural networks (ANNs) are another popular classifier used in MS data analysis. Ward et al. [13] used an ANN algorithm to analyses non-cancer and colorectal cancer samples via SELDI to identify colorectal cancer biomarkers. The ANNs with the seven highest peaks obtained 95% sensitivity and 91% specificity. Also, Chen et al. [30] used ANNs to diagnosis colorectal cancer. The proposed approach obtained 91% sensitivity and 93% specificity.

iii. Naive Bayesian (NB) Classifier

The naive Bayesian is "a simple probabilistic classifier based on the Bayesian theorem with the (naive) independence assumption" [31]. Zheng [31] compared the performance of the naïve Bayesian (NB) and the logistic regression (LR) on MS data. They found the average performance of the NB (around 90%) and the logistic regression depended on the amount of training data.

iv. Support Vector Machine (SVM) Classifier

Support vector machines (SVMs) attempt to find the best hyperplane line that separates all class A data points from class B data points. Wagner et al. [32] found that the linear SVM was the only classification method that obtained robust performance (98% accuracy). Also, Prados et al. [19] achieved 97% sensitivity and 71% specificity when used SVM-based model to classify MS data.

- **Performance Assessment of Models**

The last stage of the data mining modeling process is the assessment or validation of the model. Below, we will discuss the measures of accuracy, sensitivity, and specificity.

The classification of accuracy is calculated by "comparing the ratio of the number of correctly classified samples to the total number of samples in the test data" [33]. However, when the spread of a certain class is greater than that of other classes, the majority class will create unequal results. In this scenario, the accuracy measure will not be true. Most MS data analysis studies have used accuracy to report their results [33] [34].

There are four possible results when test decisions are built for data with two class samples: true-positive, true-negative, false-positive, and false-negative [33]. The true-positive rate is known as sensitivity. It represents the ratio of the number of correctly classified positive samples to the total number of positive samples. When the effect of incorrectly predicting a diseased person as healthy is high, high sensitivity is preferred in medical diagnoses. Specificity refers to the false-positive rate, or the probability that a healthy subject will be incorrectly classified as unhealthy [14]. When a false alarm would result in unwanted tests or treatments, high specificity is desirable here [7]. In very good classification, both sensitivity and specificity should be high, though different levels of these measures are accepted depending on the application. However, it is very hard to compare the results of different studies using only measures of sensitivity and specificity [33].Up to our knowledge, many approaches failed to achieve high accuracy. Even when high accuracy is obtained, the "black box" nature of these proposed approaches is a major issue. To address this problem and to build an accurate and understandable model, we propose to use a rule-based classifier approach along with using GAs for feature selection.

3. LITREATURE REVIEW

In this section, we explore some of the studies that use GA technique for feature selection on MS data. In addition, we illustrate some studies that use rule-based techniques as a classifier on spectrum data with classifiers.

A. Genetic Algorithm Based Feature Selection for MS Data

One popular algorithm that isused for feature selection purpose is a genetic algorithm (GA). A GA searches for optimal MS data features or biomarkers to use in the mining stage in order to distinguish patients from controls in an accurate way. Here wediscuss GA as a feature selection approach for MS data. Many studies have used GA for feature selection before applying a classifier. In 2009, Reynès et al. [35] developed a new model using a GA for feature selection and a very simple tree as a classifier. The GA in this model sought to choose a set of interesting features in a spectrum to achieve the best split points in the tree. First, the authors applied preprocessing steps to the dataset. The dataset contained 162 ovarian cancer samples and 91 control samples. Of these, 46 control samples and 81 cancer samples were randomly chosen for use as a training set; the rest (45 control and 81 cancer samples) were later used for testing. The authors obtained 98% accuracy after building the tree with three different peaks (245 Da, 434 Da, and 649 Da). The major issue in this technique when GAs return large numbers of features the DT become large and difficult to understand.

In 2004, Mohamad et al. [36] proposed a new model for applying a GA to seek and identify potential informative features using an SVM classifier. Experimental results on a breast cancer dataset (which contained 200 samples for training and 77 samples for testing) and a leukemia cancer dataset (which contained 38 samples for training and 34 samples for testing) showed the usefulness of the proposed approach for low- and high-dimension data. The authors obtained 82% accuracy for the breast cancer dataset, with 8 features, and 100% accuracy for the leukemia cancer dataset, with 50 features. In 2004, Li et al. [37] proposed a novel model used a GA for the feature selection stage and an SVM method as a classifier. The MS dataset used included 91 control samples and 162 samples from patients with ovarian cancer. Both feature selection approaches (filter and wrapper) were explored. The results showed 98% accuracy the proposed model was applied with a filter approach.

In 2002, Petricoin et al. [38] used GA for feature selection with a cluster algorithm. The proposed algorithm was applied to a training set containing 50 ovarian cancer samples and 66 control samples. The authors obtained a sensitivity of 100%, a specificity of 95%, and a rounded accuracy of 94%.In 2007 Shah and Kusiak [39] proposed a model using GA for feature selection and DT and SVM as classifiers. They applied the proposed model to three different datasets for ovarian cancer, prostate cancer, and lung cancer. The proposed model had high classification accuracy when applied to the ovarian cancer and lung cancer dataset, such that it was able to recognize the most significant features. Table1 below summarizes some of the relevant research in this field that used genetic algorithm as feature selection for mass spectrum data.

After we review some studies that using GAs for feature selection in the analysis of MS data we found that most approaches obtained a very good accuracy results. However, there are some major challenges. For example, there is no guarantee that GAs will always simultaneously find the best solution and in the same time the minimum number of discernment features. When a GA

obtains a large number of features, there will be problems using certain classifiers, such as DTs. In such cases, DTs may become very large, complex, and difficult for experts to understand. Some researchers have tried to solve this problem by adding constraints to the GA. This was the case in [39], in which the authors repeated the selection process when the number of selected features was more than 100; however, this process took a long time. Moreover, in [35], the authors added a constant to the fitness function to help it select the fewest number of features possible. However, the constant did not always work in obtaining a minimal number of features.

Table.1. Some of the research using GAs as features selection for analysis MS data

Author(s)	Year	disease	Data	Feature Selection Mothed	Data Mining Algorithm	Result
Reynès et al. [35]	2009	Ovarian Cancer	▪ 253 ovarian cancer serum samples. ▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients.	Genetic Algorithm	DT	98% Accuracy
Mohamad et al. [36]	2004	Breast Cancer	▪ 200 training samples and 77 test samples.		SVM	82% Accuracy.
		Leukemia Cancer	▪ 38 training samples and 34 test samples.			100% Accuracy.
Li et al. [37]	2004	Ovarian Cancer	▪ 253 ovarian cancer serum samples. ▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients.		SVM	98% Accuracy
Petricoin et al. [38]	2002	Ovarian Cancer	▪ 216 ovarian cancer serum samples. ▪ 100 training samples and 116 test samples.		Cluster	94% Accuracy

Shah and Kusiak [39]	2007	Ovarian Cancer	▪ 253 serum samples. ▪ 135 training samples and 118 test samples.		DT and SVM	DT: 94.07% Accuracy SVM: 97.46% Accuracy.
		Prostate Cancer	▪ 136 serum samples. ▪ 102 training samples and 34 test samples.			DT: 55.88% Accuracy SVM: 67.65% Accuracy.
		Lung Cancer	▪ 181 serum samples. 32 training samples and 149 test samples.			DT: 81.88% Accuracy SVM: 98.66% Accuracy.

After excluding all 100% accurate results due to the high chance of over-fitting, we found that the best accuracy achieved was 98.66%, which was obtained by the SVM classifier. Thus, we seek to obtain a better accuracy than this one, while simultaneously building a classifier that is easy to understand. We propose the use of a rule-based classifier, which can be understandable even when GAs return large numbers of features. This is because a rules-based classifier is easier to understandable than a DT, especially with higher numbers of features. Finally, we also seek to obtain higher classifier accuracy than that achieved by the SVM.

B. Rule-Based Classifier models for MS Data

Several machine-learning classifiers, such as DTs, SVMs, and K-nearest neighbor classifiers, have been used to successfully classify MS data. These have all achieved high predictive accuracy. However, the black-box nature of these classifiers presents major issues for developers [40] [41]. By contrast, the IF-THEN rule-based classifier can obtain satisfactory predictive accuracy, while also being easier to describe and interpret by humans than other classifiers, due to its readable IF-THEN rule structure [42]. The challenge is the extraction of a small, accurate and easy-to-interpret sets of IF-THEN rules from high-dimensional MS data. In the following, we will review various studies that have used IF-THEN rule classifiers to classify of MS data. We will then discuss these papers in order to provide a simple introduction for the development of this type of classifier.

In 2006, Ressom et al. [41] proposed a novel classifier for classifying MS data using a fuzzy IF-THEN rule-based structure. For feature selection, the authors used ant colony optimization

(ACO) with an SVM. They hoped that the combination of these two methods in the feature selection step would improve the quality of the potential biomarker identification and build an accurate fuzzy IF-THEN rules classifier. The authors collected 150 serum samples of hepatocellular carcinoma (HCC) diseases that were taken from Egypt between 2000 and 2002. Of these, 78 samples were taken from patients with HCC, and 72 samples were taken from normal individuals. After they preprocessed the samples, the authors selected 100 samples randomly as a training set, including 50 samples from the HCC patients and 50 samples from the healthy individuals. The remaining samples (28 from the HCC patients and 22 from healthy individuals) were used as a testing set for performance evaluation. The authors applied a combination of ACO and SVM to extract useful biomarkers in the feature selection stage. They found six m/z candidate biomarkers, as follows: 1863.4-1871.3, 2528.7- 2535.5, 933.6-938.2, 1737.1-1744.6, 4085.6-4097.9, and 1378.9-1381.2 Da. These six m/z candidate biomarkers were used as inputs to the IF-THEN rules classifier. The prediction accuracy of this classifier was estimated using a four-fold cross-validation method. Then, the authors used the ACO algorithm to select four rules from among the 4095 candidate rules extracted from the training dataset with the candidate biomarkers. The IF-THEN rules distinguished HCC patients from controls in the testing dataset with 91% sensitivity and 89% specificity.

Assareh and Moradi [43] proposed a model that used a t-test to select the best features and a IF-THEN rules classifier to classify the MS datasets. The dataset was for ovarian cancer, and it was made available to the public through the American National Cancer Institute (NCI) website. The ovarian cancer dataset contained 253 samples, of which 91 samples came from healthy individuals and 162 came from ovarian cancers patients. Before addressing these datasets, the authors used preprocessing to clean the datasets to enhance the classifier's performance. They binned all of the M/Z points as candidate biomarkers and applied a t-test to select the best candidate biomarkers. The t-test eliminated the biomarkers that were locally correlated, since these could correspond to the same peptide. The authors found three m/z candidate biomarkers. The proposed method achieved acceptable accuracy (86.93%) compared to two classification methods: LDA (74.24%) and KNN (68.18%).

In 2011, Wang and Palade [44] proposed a new Multi-Objective Evolutionary Algorithms-based Interpretable Fuzzy (MOEAIF) model. This model used Fuzzy C-Mean Clustering-based Enhanced Gene Selection (FCCEGS) for feature selection with fuzzy IF-THEN rules to analyze high-dimensional data, such as microarray gene expressions and MS data. The proposed model was evaluated on proteomics mass spectroscopy data from an ovarian cancer dataset containing 253 samples (91 from healthy individuals and 162 from ovarian cancer patients). Some preprocessing steps were applied to the dataset. The authors extracted eight fuzzy IF-THEN rules from the dataset (average rule length of two) using six candidate biomarkers. The candidate biomarker MZ6880.2 and the feature MZ18871.5 played important roles in most of the rules. This proposed MOEAIF model achieved 63.75% accuracy. Table 2 below summarizes some of the relevant research in this field.

In reviewing the various research papers using rule-based classifier to analyze MS data, we found that the research related to these rule-based classifiers was still very active. Various researchers had tried to improve the black-box problem of most classifiers while simultaneously achieving high predictive accuracy. Each paper proposed a model for obtaining a certain number of IF-THEN rules that would be easy for experts to understand and manipulate. However, the major challenge is improving rule accuracy by finding the best set of features. Several authors have

attempted to use different feature selection methods; however, up to our knowledge, none has achieved a higher classification accuracy.

Table 2. Research using IF-THEN rules as classifiers for the analysis of MS data.

Author(s)	Year	Diseases	Data	Features Selection Method	Data Mining Algorithm	Result
Ressom et al. [41]	2006	HCC	▪ 150 serum samples of HCC diseases ▪ 78 samples from patients with HCC and 72 samples from healthy patients	ACO-SVM algorithm	IF-THEN Rule-Based	91% sensitivity and 89% specificity
Assareh and Moradi [43]	2007	Ovarian cancer	▪ 253 ovarian cancer serum samples. ▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients	T- test		86.93% Accuracy
Wang and Palade [44]	2011	Ovarian cancer	▪ 253 ovarian cancer serum samples. ▪ 162 samples from patients with ovarian cancers and 91 samples from normal.	Fuzzy C-Mean Clustering based Enhanced Gene Selection method (FCCEGS)		63.75 % Accuracy

4. GENETIC-RULE-BASED CLASSIFIER MODEL FOR MS DATA (GRC-MS): A PROPOSED APPROACH

Given MS datasets of any diseases, the GRC-MSmodel has the following input and output:

Input: MS data obtain from controls (healthy individuals) and patients.

Output: A set of rules expressed as: I⇒C, where I refers to a set of features or biomarkers and C refers to a class label (i.e., healthy or patient).

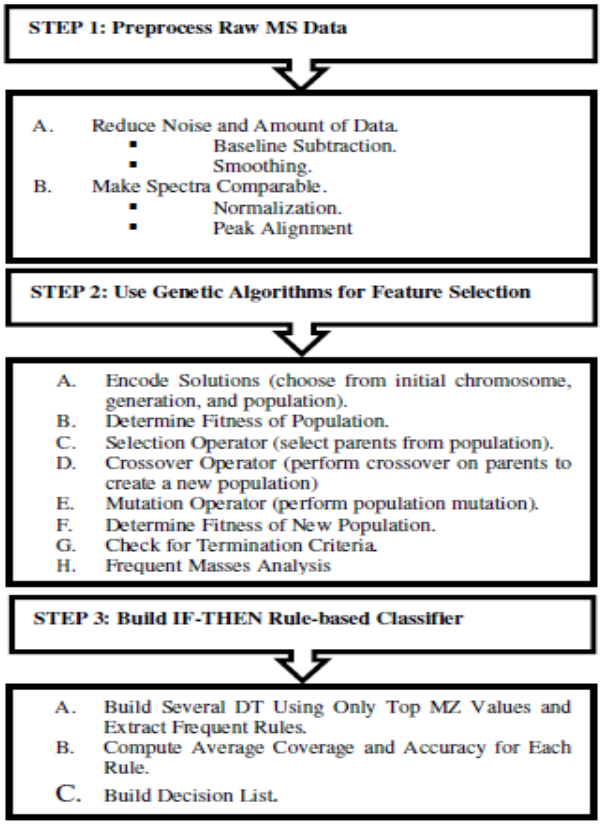


Figure 3. Steps of the GRC-MS model

The steps of the GRC-MSmodel are shown in Figure 3. The details of each step are explored in the following subsections:

STEP 1: Preprocess Raw MS Data

Each point on a spectrum is represented by two measurements: m/z and the intensity value. Sometimes, these points are affected or distorted by noise. Thus, preprocessing is needed to clean the MS data of noise and contaminants [9]. In addition, the preprocessing step must reduce or decrease the dimensions of the spectrum; this is important later for obtaining an efficient algorithm [33]. In this model, to correct the m/z and intensity values, we use the following steps: (A) Reduce Noise and Amount of Data and (B) Make Spectra Comparable.

A. Reduce Noise and Amount of Data

To remove a chemical noise baseline from a spectrum without harming the data is a challenging problem, since the wrong baseline correction may damage the spectrum, resulting in the wrong peak shape, peak position, or peak width [10]. We will use a function to estimate a low-frequency baseline. Then, we will subtract this baseline from the spectrum. Figures 4 show how the function corrects the baseline. These examples were taken from real dataset (ovarian cancer dataset) before and after the baseline's removal.

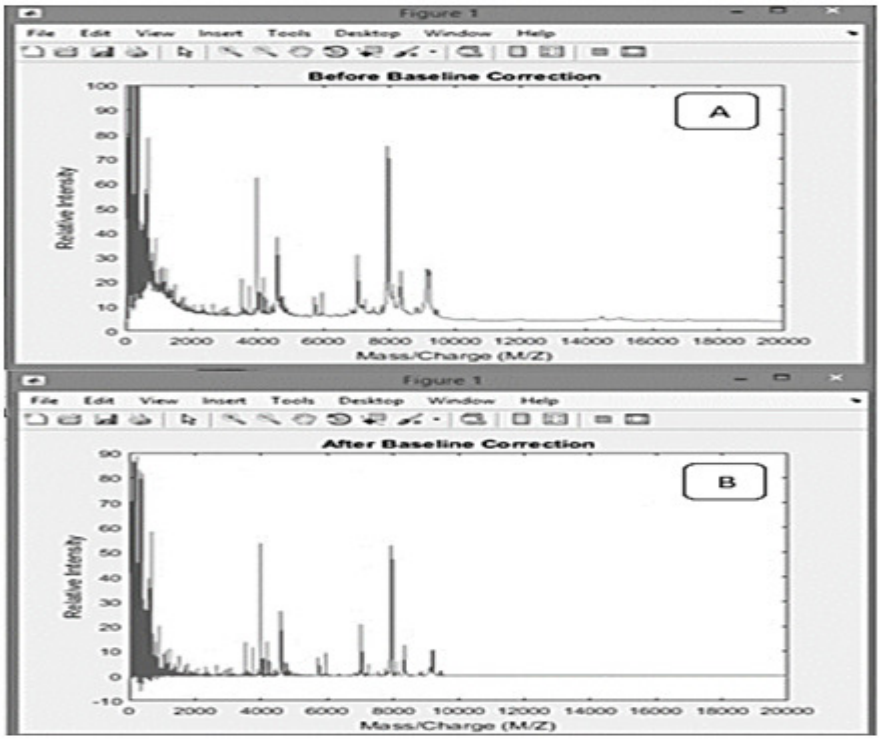


Figure 4. (A)Before baseline correction and (B) After baseline correction.

To remove electrical noise, it is important to know that spectra usually contain a combination of noises and signals. Thus, a spectrum must be de-noised to improve the validity and precision of the observed m/z values of the spectrum peaks. To accomplish this, we use Lowess smoothing and polynomial filters.

B. Make Spectra Comparable

Normalization of spectra is needed to make MS data independent of experimental differences. Normalization enables us to compare different samples, since the peak values of different spectrum fractions may be incomparable [19]. In this model, we will use the direct normalization function to calculate a re-scaled intensity value as needed. In addition, to make spectra comparable, peak alignment determines which peaks from the different spectra samples correspond to the same peak. For this, we use a sample alignment function that allows us to use a dynamic programming algorithm to assign the observed peaks in each spectrogram to the common mass/charge reference vector, if needed.

STEP 2: Use Genetic Algorithms for Features Selection

After the data preprocessing, we implement a feature selection stage, which seeks to achieve better understanding of the important features of the MS data in order to improve the classification phase later. In our model, we use GAs, which try to find optimal search solutions for problems with large datasets. Running on MS data, GAs attempt to find small sets of biomarkers that separate patient cases from control cases. This set of biomarkers, or features, is called a chromosome, such that every biomarker corresponds to a biological sample's

measurements at a given m/z value (Masse). Each chromosome is evaluate by a fitness function that attempts to find the best chromosome (set of biomarkers) for separating patients from controls. GA follows the steps outlined below:

A. Encoding Solutions (choose from initial chromosome, generation, and population)

Each "chromosome" (i.e., mathematical entity, not biological) consists of d different biomarkers or features (called genes) that are initially randomly selected from all features (since most studies used all of the MS data as features after the preprocessing steps).

B. Determine Fitness of Population

The fitness value of each chromosome is determined by the chromosome's ability to classify the training set samples into patient and control groups. In our model, we will use to compute fitness values:

Fitness value = a posteriori probability + Error rate of a linear classifier.

Note: Repeat from Step C to Step G until terminated.

C. Selection Operator (select parents from population)

The chromosome with the best fitness value is entered into the next generation, and the remaining positions are filled according to the relative fitness of the chromosomes in the parent generation (probabilistically). There are many methods for selecting the best chromosomes; we are use the roulette wheel selection method, in which the parents are selected according to their fitness. Chromosomes with greater fitness will be selected more times. Thus, the better a chromosome's fitness score, the greater its chances of being selected will be.

D. Crossover Operator (perform crossover on parents to create the new population)

The crossover can be applied to either single or double points. Each gene has an equal chance of coming from either parent. Our model use single-point and fraction crossovers to determine the fraction of the next generation population created by the crossover function.

E. Mutation Operator (perform mutation of population)

When a chromosome is chosen for transmission to the next generation, a small number of genes are randomly selected for mutation (with probabilities between 0 and 1). Once the number of genes in the chromosome to be mutated has been determined, these genes are randomly selected and replaced with genes that are not already in the chromosome. In our model, we use uniform mutation.

F. Determine the Fitness of the New Population.

G. Check for Termination Criteria.

The process is terminate when a stopping criterion, such as a specific number of high chromosomes, a maximum number of generations, or a fitness value of 100%, is obtained. We

use the average relative change in the best fitness function value over generations is less than or equal certain value or maximum number of generations is reached.

H. Frequent Masses Analysis

Frequency with which masses were select is then analyze. Then, using different number of masses form top frequency masses many times to dement best number of masses set which gives best rules accuracy.

STEP 3: Build an IF-THEN Rule-based Classifier

The IF-THEN rule-based classifier is built from training data using only the top selected features. Then, the IF-THEN rule-based classifier is used to predict the class label (i.e., healthy or patient) for the MS test data. The IF-THEN classification rule is as follows:

R: IF condition (C), THEN Class (C).
Example.

R1: "If biomarker 1 is less than threshold 1 and biomarker 2 is greater than threshold 2 and biomarker 3 is less than threshold 3, then the sample belongs to the patient group."

R2 is "If biomarker 1 is greater than threshold 1 and biomarker 2 is less than threshold 2, then the sample belongs to the healthy group."

- The LHS represent the rule condition; it is a conjunction of feature tests (biomarkers).
- The RHS denotes the rule consequent or the class label (healthy or patient).

In our work, we will build an IF-THEN rule-based classifier from a DT. In comparing the IF-THEN rule-based classifier with the decision tree, we found that the IF-THEN rule-based classifier was easier for humans to understand, especially when the DT was very huge. Then, we will assess each IF-THEN rule using rule coverage and accuracy. The Rule Ordering Scheme (i.e., Rule-Based Ordering) will then be apply. In this scheme, rule priorities are determined beforehand, and a decision list is built. This list is order according to rule quality (accuracy and coverage). The match rule that appears at the beginning of the list has the highest priority. In the event that no rule is satisfied by X, a default rule will be define for a default class, based on the training set. This class then becomes the majority class of the MS sample, encompassing all instances that are not cover by rules.

5. CASE STUDY AND RESULTS

In order to test and evaluate the accuracy of oGRC-MS model and to ensure that its rules are understandable, we apply the GRC-MS model to real data using MATLAB® software.

5.1 Dataset

We rely on open-source an MS dataset of ovarian cancer that is available to public through the clinical proteomics program of the National Cancer Institute (NCI) website(http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp). This dataset is labeled "Ovarian 8-7-02". The WCX2 protein chip was used to produce this dataset. To generate the spectrum from the samples, the

upgraded PBSII SELDI-TOF mass spectrometer was used. The dataset includes 162 ovarian cancer patients and 91 control (healthy) patients. The produced spectrum can be represented by a curved shape, in which the x-axis shows the m/z ratio (the ratio of the weight of a molecule to its charge) and the y-axis represents the intensity of the same molecule as a measure of the amount of that molecule. These datasets include peak intensity measurements at 15,154 points, as defined by the corresponding m/z values in the range of 0 to 20,000 Da.

5.2 Experimental Setup and Results

The following steps areapply by the GRC-MSmodel to the previous dataset:

1) Import MS data (raw data), using the xlsread or importdata function to load the data from an Excel® file. In Excel, the data are represented as discrete values, such that the rows show the m/z ratios and the columns represent the samples. The cells (the intersections of rows and columns) represented each molecule's intensity as a measure of the amount of that molecule in the sample. After this step is finished, we have two variables loaded into MATLAB (MZ and Y). MZ is the mass/charge vector, while Y is the intensity value for all 216 patients (control and cancer).

2) Preprocess the MS data to remove all forms of noise and all artifacts introduced to the data by applying the following functions in the following order:
• msbackadj function.
• mslowess function.
• mssgolay function.
• msnorm function.

In addition, a grouping vector is created including the type of each spectrogram and the indexing vector. This "labelling" will aid in any further analysis on this dataset.

3) Run Genetic Algorithm.

a)Create a Fitness Function for the Genetic Algorithm. In our case, the genetic algorithm tests small subsets of m/z values using the fitness function and then determines which m/z values to pass on to or remove from subsequent generations. The fitness function (biogafit) is passed to the genetic algorithm solver using a function handle. It maximizes the reparability of two classes using a linear combination of a posteriori probabilities and linear classifier error rates.
Fitness value = a posteriori probability + Error rate of a linear classifier

b) Set Genetic Algorithm Options. The GA function uses an options structure to store the algorithm parameters used to perform minimizations with the GAs. The gaoptimset function creates this options structure. The parameter values set for the GA are as follows:

- Population size: [50 100 150 200].
- Maximum number of generations: [50 100 150 200].
- Number of features: [1-10].
- Probability of crossover: [0.5 0.6 0.7 0.8].
- Probability of mutation: [0.02 0.05 0.1].
- @selectionroulette.
- @crossoversinglepoint.
- @mutationuniform.

c) Run GA to Find the Best Discriminative Features. We using the (ga) to start the GA function to decide the best feature values. We run the GA function with different times for all cases as a filter selection approach with DT. Then, we compute the DT correction rate (accuracy), the DT error rate, the DT sensitivity, and the DT specificity using 10-fold cross-validation. We also compute run time. Then, we compare the results to choice best accuracy trees. Table 3 lists the best GAs result along with their parameters. For example, in the first line we achieve 99.2 accuracy when using 200 population size, 50 generations, 0.7 crossover rate, 0.02 mutation Rate and only uses two features.The best results appears at (Table 3).

Table 3. Best GA results.

PopulationSize	Generations	No_Features	Crossover Rate	Mutation Rate	DT_CorrectRate	DT_ErrorRate	DT_Sensitivity	DT_Specifisity	Time
200	50	2	0.7	0.02	0.992	0.008	1	0.978	105.412
150	100	5	0.7	0.05	0.992	0.008	1	0.978	79.906
150	150	5	0.7	0.05	0.992	0.008	1	0.978	77.689
150	200	5	0.7	0.05	0.992	0.008	1	0.978	79.093
150	150	8	0.7	0.02	0.992	0.008	0.993	0.978	90.376
150	200	8	0.7	0.02	0.992	0.008	0.993	0.978	88.52
50	50	8	0.7	0.1	0.992	0.008	0.993	0.978	21.814
50	100	8	0.7	0.1	0.992	0.008	0.993	0.978	25.575
50	150	8	0.7	0.1	0.992	0.008	0.993	0.978	25.31
50	200	8	0.7	0.1	0.992	0.008	0.993	0.978	25.143
100	100	8	0.8	0.1	0.992	0.008	0.993	0.978	63.898
100	150	8	0.8	0.1	0.992	0.008	0.993	0.978	63.675
100	200	8	0.8	0.1	0.992	0.008	0.993	0.978	64.1
50	100	9	0.8	0.02	0.992	0.008	0.993	0.978	30.629
50	150	9	0.8	0.02	0.992	0.008	0.993	0.978	30.34
50	200	9	0.8	0.02	0.992	0.008	0.993	0.978	30.443
50	50	9	0.8	0.1	0.992	0.008	1	0.978	23.469
150	50	9	0.7	0.05	0.992	0.008	0.993	0.978	70.296

4) Frequent Masses Analysis

Using the parameters in the previous table (Table 3), we obtain 42 different masses that give us the best accuracy results. Figure 6 shows the analysis of the frequencies with which the masses are selected, where mass 8073.585 and 244.3685 appear 10 times giving the best accuracy result.

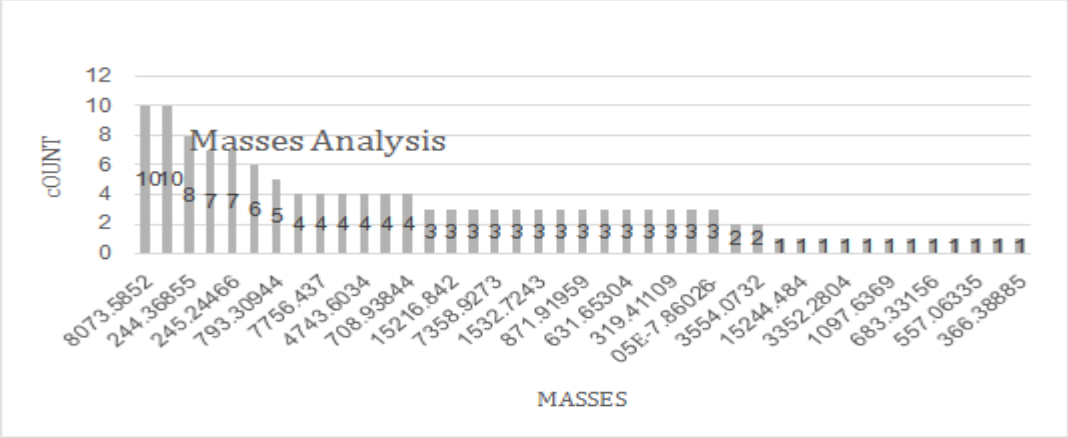


Figure 6. Masses analysis

5) Build multiple DTs From the Training Dataset Using Only Top MZ Values and Extract Frequent Rules.

This process involves built multiple DTs using different number of masses from the top frequency masses every times. Then, determine the most frequent rule in these trees for each number of masses. Steps below from A to J explain this process for each different number of masses from top two to top ten.For example**,** Using only the top two m/z values to build multiple DTs (x1=8073.5852 m/z and x2= 437.0239 m/z), which are the values extracted as top two features from the previous step. Then extract frequent rules from the multiple DTs that built by using the training dataset and compute the average coverage and accuracy of each rule using the test dataset. Note that we apply holdout validation 100 times, randomly reserve two-thirds of the dataset for training to build multiple DTs, and extract most frequent rule. The remaining one-third of the dataset is used for testing, the average coverage and accuracy are computed for the most frequent rules every time.

- R1: IF MZ (437.0239) >= 1.22269 THEN Class = Cancer.
- R2: IF MZ (437.0239) < 1.22269 and MZ (8073.5852) < 0.29102 THEN Class = Cancer.
- R3: IF MZ (437.0239) < 1.22269 and MZ (8073.5852) >= 0.29102 THEN Class = Normal.

Last, build a decision list using accuracy values.

Table 4.Rules accuracy values

Rule	Average Coverage Percentage	Average Accuracy Percentage
R1	100	95.74
R2	100	38.35
R3	100	97.62
Overall (R1+R2+R3)	100	98.80

Table 5.Decision list using accuracy values

Priority	Rule	Class
1	R3	Normal
2	R1	Cancer
3	R2	Cancer
Other	Default	Cancer

6. DISCUSSION

The results show that the GRC-MS classifier model achieves very good results when applied to the analysis of ovarian cancer datasets with different numbers of features or masses that used in the model (Table 6). We observed that 437.0239,244.36855, 8073.5852 and 793.30944 m/z were significantly discriminative masses that can be potential biomarkers for ovarian cancer. Table 7 lists the frequently occurring masses that play important roles in most of the rules.

Table 6. The GRC-MS classifier results.

No. Features	No. Rule	Accuracy
2	3	98.8095
3	3	98.8142
4	3	99.2118
5	3	99.2447
6	4	99.5731
7	4	99.6099
8	4	99.6112
9	4	99.6016
10	4	99.7038

Table 7. Frequently occurring masses that play important roles in most of the rules

Mass	Frequency
437.0239	9
244.36855	5
8073.5852	4
793.30944	4
681.86861	1

In Table 8,shows that our GRC-MS classifier model provides highly competitive accuracy (99.7%) when compared to other existing classifier models, when applied to an ovarian cancer dataset. In addition, our model also provides highly comprehensible rules that facilitate the translation of raw data into easy-to-understand knowledge that can help experts.

Table 8. Results of some existing classifier models

Author(s), Year	Feature Selection Method	Data Mining Algorithm	Result (Accuracy)
Reynèset al. [35], 2009	GA	DT	98%
Li et al.[37], 2004	GA	SVM	98%
Petricoin et al.[38], 2002	GA	Cluster	94%
Shah and Kusiak [39], 2007	GA	DT	94.07%
Shah and Kusiak [39], 2007	GA	SVM	97.46%
Assareh and Moradi [43], 2007	T- test	IF-THEN Rule-Based	86.93%
Wang and Palade [40], 2011	Fuzzy C-Mean Clustering-based Enhanced Gene Selection Method	IF-THEN Rule-Based	63.75 %

7. CONCLUSION AND FUTURE WORK

Many studies have sought to increase the accuracy of diagnoses by analyzing MS data and finding biomarkers. Some of these studies have proposed approaches capable of high accuracy, sensitivity, and specificity, while other studies have failed to obtain satisfactory results. One major issue remains: How can an accurate model that avoids the “black box” limitation be built? The “black box” produces such problems as a lack of knowledge flow between the system and the expert. To address this problem and build a model capable of yielding accurate diagnoses that are easy for experts to understand, we used a ruled-based technique to build a classifier model to analyze MS data. Recently, significant attention has been paid to the use of rule-based classification techniques because of their unique ability to provide meaningful outcomes.

In addition, we apply a GA in the feature selection stage to increase the quality and accuracy of the p **GRC-MS** classifier model. In previous research, excellent results have been obtained through the combination of GA with different types of classifiers. In order to test the validity, accuracy, and performance of the **GRC-MS** model, we conducted an experimental study using open-source databases. In this experiment, we first applied several preprocessing steps to prepare the MS data for the **GRC-MS** model. These steps included reducing the noise in the data and the amount of data, identifying and extracting peaks, and normalizing and aligning the data. We found that the **GRC-MS** classifier model enhance the accuracy and meaningfulness of the MS data analysis results. As a future work, we aim to apply the **GRC-MS** model to another MS dataset or other high-dimension dataset, such as a microarray gene expression dataset. We also aim to develop more effective fitness functions for the GA.

REFERENCES

[1] H. Fernández, “Comparison of MALDI-TOF mass spectrometry data preprocessing techniques and their effect in sample classification.”

[2] M. Durairaj and V. Ranjani, “Data mining applications in healthcare sector a study,” International Journal of Scientific & Technology Research, vol. 2, pp. 29-35, 2013.

[3] Mass Spectrometry. (2015, Feb. 15). [Online]. Available: http://www.premierbiosoft.com/tech_notes/mass-spectrometry.html
[4] E. D. Hoffman and V. Stroobant, Mass Spectrometry: Principles and Applications. Belgium, John Wiley & Sons Ltd., 2007.
[5] P. Veltri, "Algorithms and tools for analysis and management of mass spectrometry data," Journal of Briefings in Bioinformatics, vol. 9, pp. 144-155, 2008.
[6] M. Cannataro, P. H. Guzzi, T. Mazza, G. Tradigo, and P. Veltri, "Preprocessing of mass spectrometry proteomics data on the grid," in Proc. 18th IEEE Symp. on Computer-Based Medical Systems, 2005, pp. 549-554.
[7] S. Bachmayer, "Preprocessing of mass spectrometry data in the field of proteomics," M.S. thesis, University of Helsinki, Finland, pp. 8-13, 2007.
[8] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins, and Y. Huang, "Review of peak detection algorithms in liquid-chromatography-mass spectrometry," Journal of Current Genomics, vol. 10, p. 388-401, 2009.
[9] R. Kandaa and R. Glendinning, "Mass spectrometry for environmental and wastewater monitoring," Journal of Spectroscopy Europe, vol. 23, pp. 15-27, 2011.
[10] M. Katajamaa and M. Orešič, "Data processing for mass spectrometry-based metabolomics," Journal of Chromatography A, vol. 1158, pp. 318-328, 2007.
[11] M. A. Rogers, P. Clarke, J. Noble, N. P. Munro, A. Paul, P. J. Selby, and R. E. Banks, "Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis identification of key issues affecting potential clinical utility," Journal of Cancer Research, vol. 63, pp. 6971-6983, 2003.
[12] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Journal of Bioinformatics, vol. 23, pp. 2507-2517, 2007.
[13] D. Ward et al., "Identification of serum biomarkers for colon cancer by proteomic analysis," British Journal of Cancer, vol. 94, pp. 1898-1905, 2006.
[14] E. F. Petricoin et al., "Use of proteomic patterns in serum to identify ovarian cancer," The Lancet, vol. 359, pp. 572-577, 2002.
[15] G. Ge and G. W. Wong, "Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles," Journal of BMC Bioinformatics, vol. 9, pp. 275-287, 2008.
[16] P. Yang and Z. Zhang, "A clustering based hybrid system for mass spectrometry data analysis," Journal of Pattern Recognition in Bioinformatics, pp. 98-109, 2008.
[17] C. Yang, Z. He, and W. Yu, "Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis," Journal of BMC Bioinformatics, vol. 10, pp. 4-14, 2009.
[18] "Smoothing," (2015, Feb. 22). [Online]. Available: http://www.wavemetrics.com/products/IGORPro/dataanalysis/signalprocessing/smoothing.htm.
[19] J. Prados, A. Kalousis, J. C. Sanchez, L. Allard, O. Carrette, and M. Hilario, "Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents," Journal of Proteomics, vol. 4, pp. 2320-2332, 2004.
[20] A. Thomas, G. D. Tourassi, A. S. Elmaghraby, R. Valdes Jr, and S. A. Jortani, "Data mining in proteomic mass spectrometry," Journal of Clinical Proteomics, vol. 2, pp. 13-32, 2006.
[21] T. N. Vu and K. Laukens, "Getting your peaks in line: A review of alignment methods for NMR spectral data," Journal of Metabolites, vol. 3, pp. 259-276, 2013.
[22] Y. Su et al., "Diagnosis of gastric cancer using decision tree classification of mass spectral data," Journal of Cancer Science, vol. 98, pp. 37-43, 2007.
[23] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," Artificial Intelligence in Medicine, vol. 31, pp. 91-103, 2004.
[24] R. Blanco, P. Larranaga, I. Inza, and B. Sierra, "Gene selection for cancer classification using wrapper approaches," International Journal of Pattern Recognition and Artificial Intelligence, vol. 18, pp. 1373-1390, 2004.
[25] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," BMC Bioinformatics, vol. 7, p. 3, 2006.
[26] S. Ma and J. Huang, "Regularized ROC method for disease classification and biomarker selection with microarray data," Bioinformatics, vol. 21, pp. 4356-4362, 2005.

[27] F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, and P. Veltri, “A time series approach for clustering mass spectrometry data,” Journal of Computational Science, vol. 3, pp. 344-355, 2012.
[28] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman, “Diagnosis of ovarian cancer using decision tree classification of mass spectral data,” Journal of Biomedicine and Biotechnology, vol. 5, pp. 391-404, 2003.
[29] Y. Su et al., “Diagnosis of gastric cancer using decision tree classification of mass spectral data,” Journal of Cancer Science, vol. 98, pp. 37-43, 2007.
[30] Y. D. Chen, S. Zheng, J.-K. Yu, and X. Hu, “Artificial neural networks analysis of surface-enhanced laser desorption/ionization mass spectra of serum protein pattern distinguishes colorectal cancer from healthy population,” Journal of Clinical Cancer Research, vol. 10, pp. 8380-8385, 2004.
[31] J. Zheng, “A comparison of naïve Bayes and logistic regression classifiers for mass spectrometry data,” Journal of Proteomics & Bioinformatics, 2007.
[32] M. Wagner, D. Naik, and A. Pothen, “Protocols for disease classification from mass spectrometry data,” Journal of Proteomics, vol. 3, pp. 1692-1698, 2003.
[33] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Francisco, CA, Morgan Kaufmann Publishers Inc., 2011.
[34] E. F. Petricoin et al., “Use of proteomic patterns in serum to identify ovarian cancer,” Journal of Mechanisms of Disease, vol. 359, pp. 572-577, 2002.
[35] C. Reynès, R. Sabatier, N. Molinari, and S. Lehmann, "A new genetic algorithm in proteomics: Feature selection for SELDI-TOF data," Computational Statistics & Data Analysis, vol. 52, pp. 4380-4394, 2008.
[36] M. S. Mohamad, S. Deris, S. Yatim, and M. Othman, "Feature selection method using genetic algorithm for the classification of small and high dimension data," in Proc. of the 1st Int. Symposium on Information and Communication Technology, 2004, pp. 13-16.
[37] L. Li et al., "Data mining techniques for cancer detection using serum proteomic profiling," Artificial Intelligence in Medicine, vol. 32, pp. 71-83, 2004.
[38] E. F. Petricoin et al., “Use of proteomic patterns in serum to identify ovarian cancer,” Journal of Mechanisms of Disease, vol. 359, pp. 572-577, 2002.
[39] S. Shah and A. Kusiak, "Cancer gene search with data-mining and genetic algorithms," Computers in Biology and Medicine, vol. 37, pp. 251-261, 2007.
[40] Z. Wang and V. Palade, "Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis," BMC Genomics, vol. 12, p. S5, 2011.
[41] H. W. Ressom et al., "Biomarker identification and rule extraction from mass spectral serum profiles," in proc IEEE Symp. Computational Intelligence and Bioinformatics and Computational Biology, 2006, pp. 1-7.
[42] J. M. Sorace and M. Zhan, "A data review and re-assessment of ovarian cancer serum proteomic profiling," BMC Bioinformatics, vol. 4, p. 24-35, 2003.
[43] A. Assareh and M. H. Moradi, "Extracting efficient fuzzy if-then rules from mass spectra of blood samples to early diagnosis of ovarian cancer," in Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB'07. IEEE Symposium on, 2007, pp. 502-506.
[44] Z. Wang and V. Palade, "Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis," BMC Genomics, vol. 12, p. S5, 2011.

UNDERSTANDING PEOPLE TITLE PROPERTIES TO IMPROVE INFORMATION EXTRACTION PROCESS

Saleem Abuleil and Khalid Alsamara

MMIS Department, Chicago State University, Chicago, USA

ABSTRACT

In this paper, we introduce a new approach to tackle the process of extracting information about people mentioned in the Arabic text. When a person name is mentioned in the Arabic text usually it is combined with a title, in this paper the focus is on the properties of those titles. We have identified six properties for each title with respect to gender, type, class, status, format, and entity existence. We have studied each property, identified all attributes and values that belong to each one of them and classified them accordingly. Sometimes person title is attached to an entity; we have also identified some properties for these entities and we show how they work in a harmony with person title properties. We use graphs for the implementation, nodes to represents person title, person name, entity and their properties, where edges are used to present inherited properties from parent nodes to child nodes.

KEYWORDS

People Titles, Title Properties, NLP, Arabic Language

1. INTRODUCTION

Information Extraction (IE), as defined in the Message Understanding Conferences, has been traditionally defined as the extraction of information from a text in the form of text strings and processed text strings that are placed into slots labeled to indicate the kind of information that can fill them. The problem of extracting information from a large document collection can be approached using many different algorithms. The three classic models used in information extraction, (under which all these algorithms can be loosely grouped), are called Rule-based, Pattern Learning, and Supervised Learning. Most of the Arabic Named Entity Recognition (NER) systems use keywords such as titles to tag proper name phrases in the text, once they tag proper name phrase they use either rule-based systems or statistical approach to tag proper names and extract information about them. Using titles to tag proper names in the Arabic text is an important technique that has been used widely, but titles have been used as keywords for the purpose of identifying proper name phrases and tag proper names without studying and exploring their properties. Our technique in this paper is to identify and use title properties and attributes to enhance the result of extracting information about people names in the Arabic text.

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 37–46, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80403

2. LECTURER

Al-Kouz [1] presented a framework designed for mining the explicit and implicit lexical semantic information impeded in the structure and the content of Aljazeera.net. Furthermore, it provides an efficient and structured access to the resulted semantic graph, the authors also claim in their paper that Aljazeera.net is professionally edited and has rich semantic structure and it establishes an asset, an impediment and a challenge for research in Arabic Natural Language Processing. Abdallah [2] proposed a simple method for integrating machine learning with rule-based systems and implement this proposal using the state-of-the-art rule-based system for NERA, the experimental evaluation shows that their integrated approach increases the F-measure by 8 to 14% when compared to the original (pure) rule-based system and the (pure) machine learning approach, and the improvement is statistically significant for different datasets, more importantly, their system outperforms the state-of-the-art machine-learning system in NERA over a benchmark dataset. Abdul Hamid [3] introduced simplified yet effective features that can robustly identify named entities in Arabic text without the need for morphological or syntactic analysis or gazetteers, a CRF sequence labeling model is trained on features that primarily use character n-gram of leading and trailing letters in words and word n-grams, the proposed features help overcome some of the morphological and orthographic complexities of Arabic.

Abuleil [4] presented a new technique to extract names from the text by building a database and graphs to represent the words that might form a name and the relationships between them. First, they mark the phrases that might include names, second they build graphs to represent the words in these phrases and the relationships between them, and third, they apply rules to find the names. Benajiba [6] investigated the impact of using different sets of features in three discriminative machine learning frameworks, namely, support vector machines, maximum entropy and conditional random fields for the task of named entity recognition ,they explore lexical, contextual and morphological features and nine data-sets of different genres and annotations; they measure the impact of the different features in isolation and incrementally combine them in order to evaluate the robustness to noise of each approach.

Chen [8] described their system for the CoNLL-2012 shared the task, which seeks to model co-reference in Onto Notes for English, Chinese, and Arabic; they adopt a hybrid approach to co-reference resolution, which combines the strengths of rule-based methods and learning-based methods, they official combined score over all three languages is 56.35. In particular, their score on the Chinese test set is the best among the participating teams. Habash [9] made an argument that is the many differences between Dialectal Arabic and Modern Standard Arabic (MSA) pose a challenge to the majority of Arabic natural language processing tools, which are designed for MSA, so in their paper retarget an existing state-of-the-art MSA morphological tagger to Egyptian Arabic (ARZ), their evaluation demonstrates that their ARZ morphology tagger outperforms its MSA variant on ARZ input in terms of accuracy in part-of-speech tagging, diacritization, lemmatization and tokenization; and in terms of utility for ARZ-to English statistical machine translation. Pasha [10] presented MADAMIRA, a system for morphological analysis and disambiguation of Arabic that combines some of the best aspects of two previously commonly used systems for Arabic processing, MADA (Habash and Rambow, 2005; Habash et al., 2009; Habash et al., 2013) and AMIRA (Diab et al., 2007). MADAMIRA improves upon the two systems with a more streamlined Java implementation that is more robust, portable, extensible and is faster than its ancestors by more than an order of magnitude.

3. PEOPLE TITLE PROPERTIES

When a person name is mentioned in the Arabic text usually it is attached to a title. We have studied these titles and identified different properties for each one of them; we also identified some attributes and values for each property. We have identified six properties for each title with respect to gender, type, class, status, format, and existence of an entity. In this section, we explain each one of them and we show some examples in table 1.

Gender: in the Arabiclanguage, there are two values for this property masculine and feminine. In Arabic language to form a feminine title from the masculine, you simply add "taa' marbuta" which looks like (ة, ـة) to the end of the title, for exampleوزيرWazer (he) Minister is a masculine and to form the feminine from it we add "taa' marbuta" " ة" to the end of the title وزيرةWazertn (she) Minister

Type: We have classified title into three Types:

- Occupational title that indicates a position or job of the person like Manager مدير Minister وزيرPresident رئيسand Consultant مستشار
- A social title like Mr. سيد, Ms. انسه,and Mrs. سيدة.
- Professional title that refers to a certain profession like engineerمهندس, physician طبيب, attorney محامي, and teacher استاذ

A person might hold two titles at the same time such as(Mr. and consultant), and(engineer and manager), etc. some titles could be used for two classes like الشيخ Sheikh, could be used for social or occupational.

Class: based on job field that people they hold we have classified titles into different classes: politics, religion, education, sport, media, industry, military, etc. some titles could be used for two types like الشيخ Sheikh, could be used for religion or politics, some titles like president and manager could be classified into different classes politics, education, sports, industry, etc. to identify the class for these cases we use the entities as we discuss later. Some titles do not have a class such as Mr. and Mrs.

Status: person title could be simple or compound, simple title has one word such as سيد Mr. and وزير Minister, compound title has two words like ولي العهد Crown Prince and الناطق الرسمي Official Speaker

Format: there are two formats of the title either defined or indefinite. Arabic word starts with ال (the) to define it. Al- (ال) is the definite article in the Arabic language. For example, the word وزير wazer "Minister" can be made definite by prefixing it with al-, resulting in الوزير al-wazer "the Minister". Consequently, al- is typically translated as The in English. A defined title that starts with الand the word followed is not a verb, adjective, nationality, or particle then the noun is most likely is a person name.

Entity Existence: Sometimes an entity comes between person title and person name, entity existence property has two values either Yes or No. Most likely if a person title is defined no entity to follow. More details about entities are discussed in the next section

4. ENTITIES AND THEIR PROPERTIES

Sometimes entity or nationality comes between person title and person name. We have studied hundreds of cases and based on our study we identified four properties for each one of them with respect to type, class, gender, and status. See table 2. In this paper we handle the following three most common scenarios:

- Person Title + **Entity (both title and name are mentioned)** + Person Name

  Example: رئيس**دولة تركيا**اردغان President of **Turkey State**Erdoğan

  In this example entity title (دولة State) and entity name (اردغانErdoğan) are mentioned

- Person Title + **Entity (title is omitted and name is mentioned**) + Person Name

  Example: مراسل **الجزيرة** إلياس كرام Report of **Aljazeera** Alysa Karram

  In this example entity tile (اخباريهقناة News Channel) is omitted and entity name (**الجزيرة Aljazeera)** is mentioned

- Person Title + **Entity (title is mentioned and name/nationality is omitted)** + Person Name

  Example: الناطق الرسمي باسم **الحكومه**الدكتور محمد المومني

  Official speaker of the **government** Dr. Muhammad Almumani

  In this example entity title (الحكوكة government) is mentioned and nationality (الاردنيه Jordanian) is omitted

In some cases it is difficult to identify the class property of person title like President رئيس and Managerمدير where each one of them can be classified into different classes such as politic, sport, industry, but when an entity is maintained, and by using the value of the class of that entity, it helps to identify the class of the person title, for example when a company name is attached to person title رئيس President, the title is classified as industry category and when university name is attached to the same person tile رئيس president, the title this time is classified as education category.

In this paper, we identify nationality as an adjective and we identify four properties for it with respect to gender, type, format and country. Gender could be either masculine or feminine, the type has the value nationality, and the country has the value which country the nationality belongs to. When nationality comes directly after theرئيس president, we add a class property with a value politics.

Table 1.Title Properties

Gender		Format	Type	Class	Entity Existence
Masculine	Feminine				
وزير Minister	وزيرة	Indefinite	Occupational	Politics	Yes
الوزير The Minister	الوزيرة	Defined	Social	Politics	No
لاعب Player	لاعبة	Indefinite	Occupational	Sport	Yes
الاعب The Player	الاعبة	Defined	Occupational	Sport	No
ولي العهد Crown Prince	وليه العهد	Indefinite	Occupational	Politics	No
الناطق الرسمي Official Speaker	الناطقة الرسمية	Defined	Occupational	Follows entityclass	Yes
مدير Manager	مديره	Indefinite	Occupationally	Follows entityclass	Yes
المدير The Manager	المديرة	Defined	Social	Follows entityclass	No
المحامي The Attorney	المحاميه	Defined	Professional	Law	No
الشيخ The Sheikh	الشيخه	Defined	Social Occupational	Politic or Religion	No
السيد Mr.	السيدة Mrs.	Defined	Social	N/A	No
المدرس The Teacher	المدرسة	Indefinite	Professional	Education	No
مراسل Reporter	مراسلة	Indefinite	Media	Media	Yes

Table 2.Entity Properties

Title	Gender	Class	Type
دولة state	Feminine	Politics	Location
حكومة	Feminine	Politics	Location
مملكة kingdom	Feminine	Politics	Location
وزارة ministry	Feminine	Politics	Organization
جمعية society	Feminine	Social	Organization
نادي club	Masculine	Social / Sport	Organization
جامعة university	Feminine	Education	Organization
كلية college	Feminine	Education	Organization
مدرسة school	Feminine	Education	Organization
ملعب stadium	Masculine	Sport	Location
قناه اخبارية News Channel	Feminine	Media	Organization
جامع mosque	Masculine	Religion	Organization
كنيسة church	Feminine	Religion	Organization

5. ANALYSIS

We use graphs to implement the concepts in this paper; we use nodes to represents person title, person name, nationality, entity and their properties, and we use edges to present inherited properties from parent node to child node, see fig 1. We tag person name phrases in the text, each phrase starts with a title and terminated with a person name, next we tag the elements in the tagged phrases with respect to titles, entities, and nationalities and then apply the concepts of this paper to identify the properties of each one of them. Properties are inherited from one node to another (parent to child) and once a child gets inherited property from the parent they also forward this information to next node in the graph. When reach the last node in the graph we process all information from all nodes and produce the results. Harmony of inherited information between nodes is also validated with respect to the properties of the titles, entities, and nationalities. There is should be a match between two titles belong to the same person, same gender and format values between adjective (nationality) and the element presided to it either a title or an entity

In figure 2 we illustrate different scenarios about the same person and we show how the person inherits all properties from all parent nodes, the person name is Albright, Albright is a female, she is a politician, and her occupation is the minister of the USA foreign ministry. Example 1 contains two titles and one adjective (nationality), the first title and the adjective both have the same values of format property and gender property, both titles mentioned for the same person and they have the same value of gender property. Example 3contains one entity which is الخارجيـــه foreign and one adjective الامريكيـــه American, since both of them have the same value of format property "defined", the title وزيـرة minister has a different format property value "indefinite", then the nationality refers to the organization and not to the person title, the entity title وزارة ministry is omitted, both titles and the nationality of this example belong to the same person they have the same value of gender property.

<table>
<tr><td colspan="4">Example 1</td></tr>
<tr><td>الوزيرة The Minister</td><td>Title</td><td rowspan="2">Defined</td><td rowspan="2">Feminine*</td></tr>
<tr><td>الامريكية American</td><td>Adjective</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td>السيدة The Mrs.</td><td>Title</td><td colspan="2">Feminine*</td></tr>
</table>

<table>
<tr><td colspan="4">Example 3</td></tr>
<tr><td>وزيرة Minister</td><td>Title</td><td>Indefinite</td><td rowspan="3">Feminine*</td></tr>
<tr><td>الخارجية Foreign</td><td>Entity</td><td rowspan="2">Defined</td></tr>
<tr><td>الامريكية American</td><td>Adjective</td></tr>
<tr><td colspan="4"></td></tr>
<tr><td>السيدة The Mrs.</td><td>Title</td><td colspan="2">Feminine*</td></tr>
</table>

In figure 3 we show different scenarios where the value of the class property of person title is uncertain and we are going to use the value of the class property of the entity to identify it. In example 1, the value of the class property of the entity is education, and then we use it for the value of the class property of the title "رئيس". Example 2, the value of the class property of the entity is politics, and then the value of the class property of the title "رئيس" is also politics, the second title is Sheik "شيخ" and could have two possibilities either politics or religion but since the first title has the class property value is politics we select the value politics as well for the second

title. In example 3, the value of the class property of the entity is industry, and then the value of the class property of the title "رئيس" is also industry.

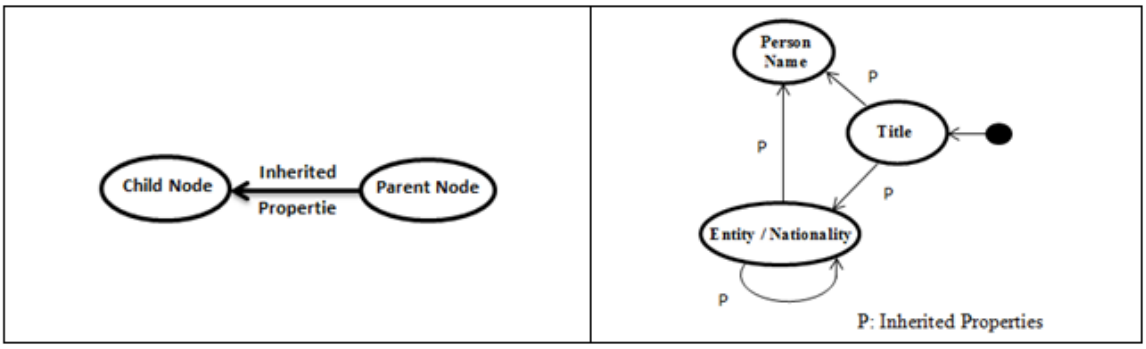
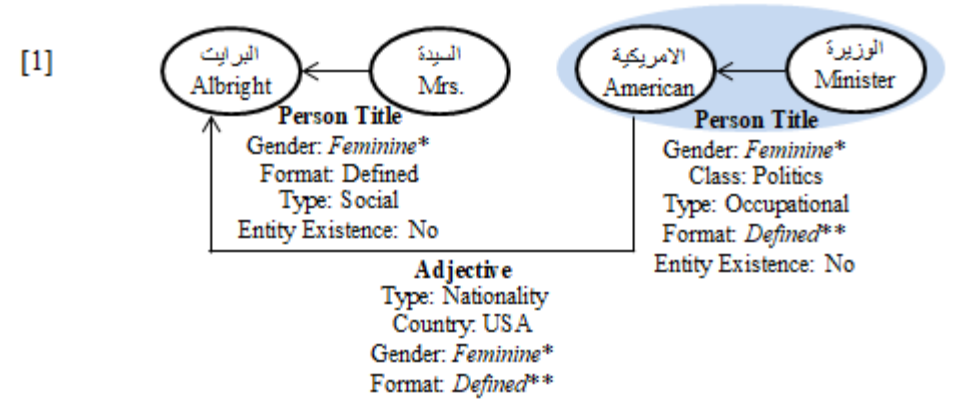
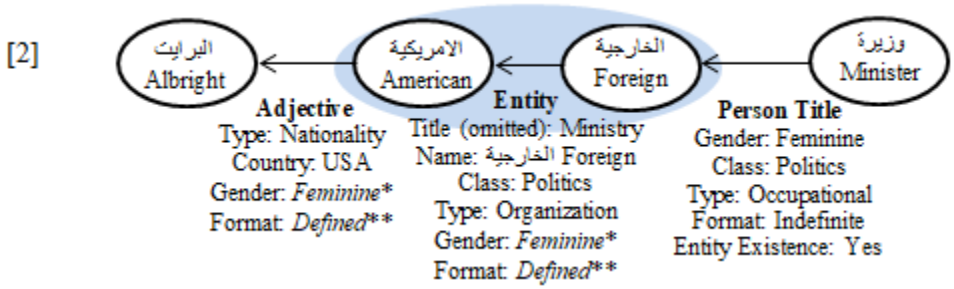


Fig 1. Graph

6. CONCLUSION

In this paper we have studied person titles, entities and nationalities attached to them, identify the properties for each one of them and used this information to extract information about people mentioned in the Arabic text we also validated the inherited property values between the nodes in the graph. Our source of data is Ajazeera.net, further analysis to be done in the future.





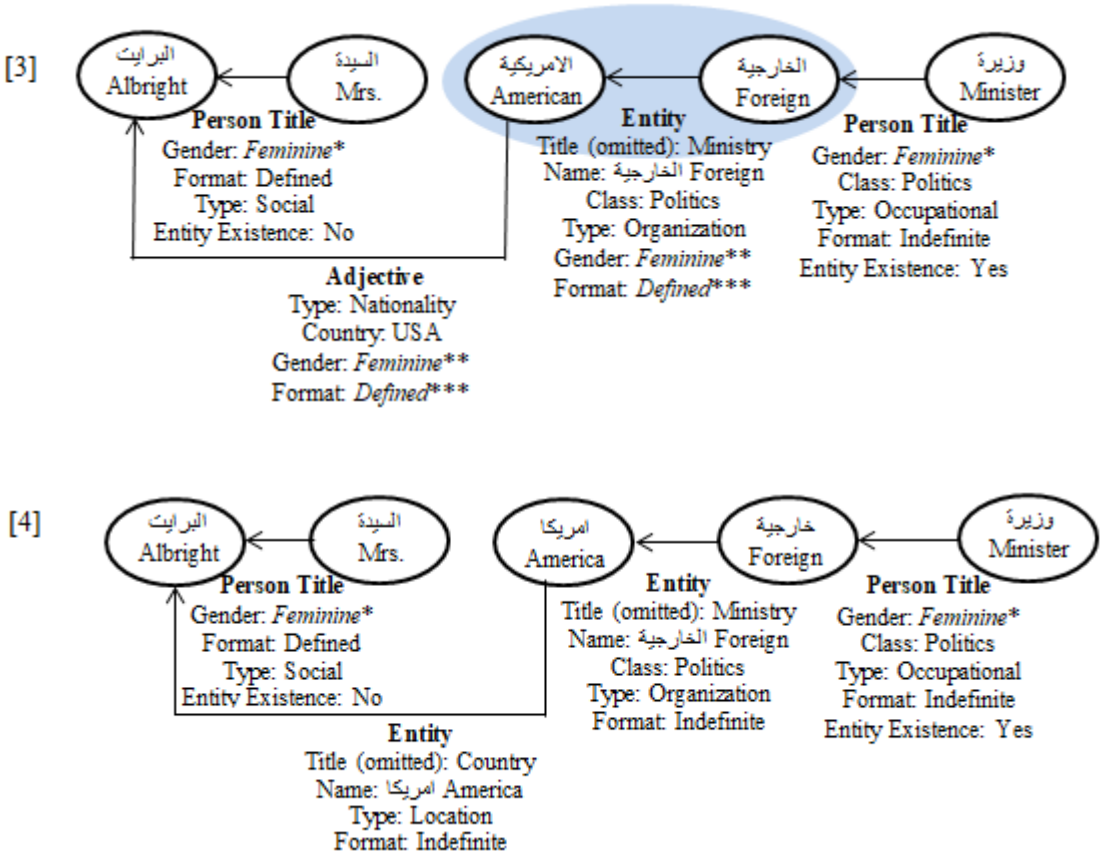
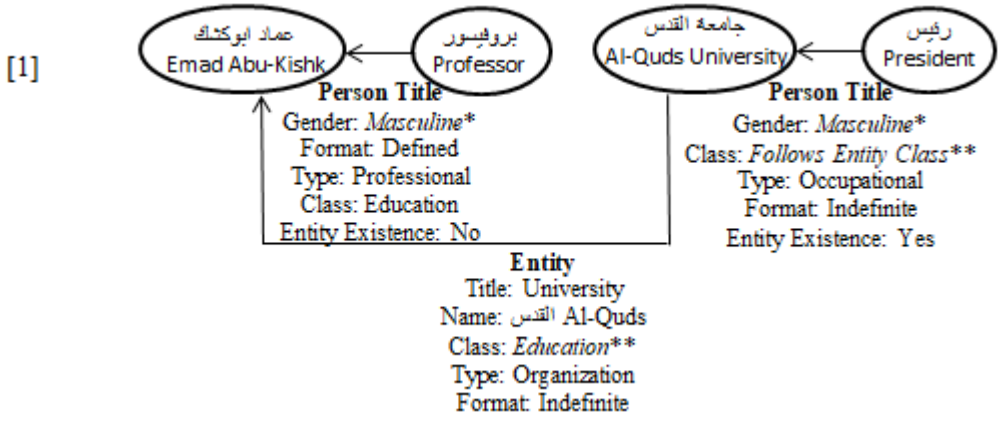


Fig 2.Illustration A



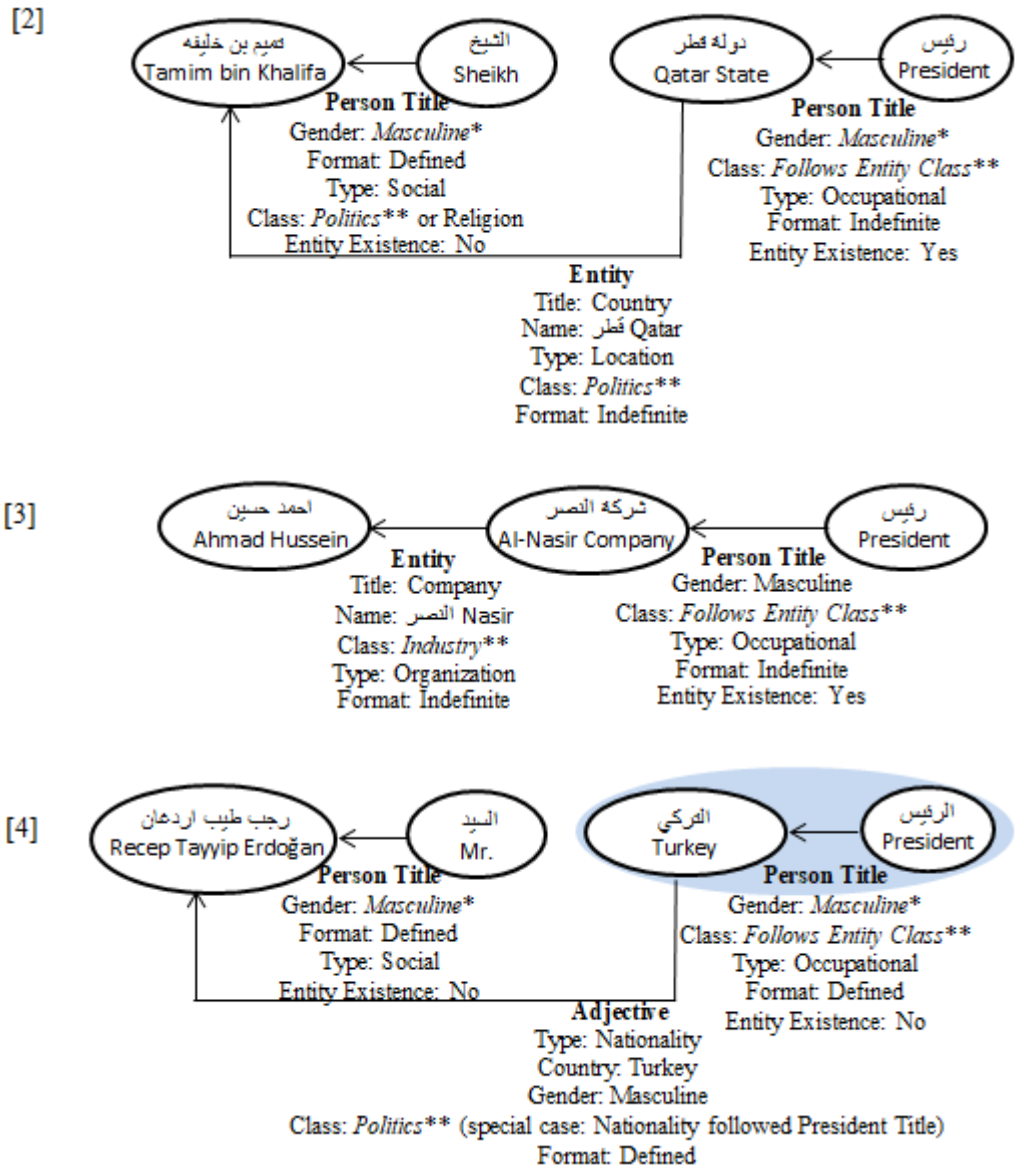


Fig 3. Illustration B

REFERENCES

[1] Al-Kouz, A., Awajan, A., Jeet, M., Al-Zaqqa, A.: Extracting Arabic semantic graph from Aljazeera. net. In: 2013 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT), (pp. 1–6). IEEE, Dec 2013

[2] Abdallah, S., Shaalan, K.,Shoaib, M.: Integrating rule-based system with classification for Arabic named entity recognition. In: Gelbukh, A. (ed.) CICLing 2012. LNCS, vol. 7181, pp. 311–322. Springer, Heidelberg (2012)

[3] Abdul Hamid, A., Darwish, K.: Simplified feature set for Arabic named entity recognition. In: Proceedings of the 2010 Named Entities Workshop, pp. 110–115. Association for Computational Linguistics, Uppsala (2010)

[4] Abuleil, S.: Extracting names from Arabic text for question-answering systems. In: Pro-ceedings of Coupling approaches, coupling media and coupling languages for information retrieval, RIAO 2004,pp. 638-647. Avignon(2004)

[5] Abuleil, S.: Hybrid system for extracting and classifying Arabic proper names. In: Pro-ceedings of the fifth WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, AIKED 2006, pp. 205-210. Madrid (2006)

[6] Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition: A feature-driven study. IEEE Trans. Audio Speech Lang. Process. 17(5), 926–934 (2009)CrossRef Google Scholar

[7] Benajiba, Y., Diab, M., Rosso, P.: Arabic named entity recognition: An svm-based ap-proach. In: The International Arab Conference on Information Technology, ACIT 2008 (2008)

[8] Chen, C., Ng, V.: Combining the best of two worlds: a hybrid approach to multilingual coreference resolution. In: Joint Conference on EMNLP and CoNLL-Shared Task, pp. 56–63. Association for Computational Linguistics, July 2012Google Scholar

[9] Habash, N., Roth, R., Rambow, O., Eskander, R., Tomeh, N.: Morphological analysis and disambiguation for dialectal Arabic. In: HLT-NAACL, pp. 426–432 (2013)Google Scholar

[10] Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Roth, R.M.: (2014). Madamira: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In: Proceedings of the Language Resources and Evaluation Conference (LREC). Reykjavik, Iceland Google Scholar

STRATEGY OF THE REMOVE AND EASY TBT IN GCC 6 COUNTRIES

Yong-Jae Kim

Department of Business Administration, Korea Polytechnic University,
Prof. ph.D

ABSTRACT

The last technical barriers to trade(TBT) between countries are Non-Tariff Barriers(NTBs), meaning all trade barriers are possible other than Tariff Barriers. And the most typical examples are (TBT), which refer to measure Technical Regulation, Standards, Procedure for Conformity Assessment, Test & Certification etc. Therefore, in order to eliminate TBT, WTO has made all membership countries automatically enter into an agreement on TBT. In this study, the elimination strategy of TBT with aid of technical regulations or standards is excluded, and only the conformity assessment shall be considered as the strategic measure of eliminating TBT in GCC(Gulf Cooperation Council) 6 countries. The measure for every membership country to accord with the international standards corresponding to their technical regulations and standards, is only to present TBT related Specific Trade Concern(STC) to WTO. However, each of countries retains its own conformity assessment area, and measures to settle such differences are various as well. Therefore, it is likely required an appropriate level of harmonization in them to carry forward this scheme. KTC(Korea Testing Certification) written MRA with GCC test & certification company in 2015 years. So Korea exporting company can export to GCC goods with attached test & certification documents in Korea. To conclude, it is suggest MRA for the remove and reduce TBT to increase export and import among countries.

KEYWORDS

FTA, Standards, Conformity Assessment, TBT(Technical barrier to trade), MRA, WTO

1. INTRODUCTION

This paper purpose make to remove and to easy TBT of industrial products such as IT, S/W, IOT, BigData, Home network. Research methodology is review 2nd data analysis and focus group Interview Government officer, Professor and CEO.

This paper compare & analyze International rule & system as follow.

First, It is to compare & analyze the standard, technical regulation, Test & certification procedure and Inspection. Second, it is review electric/electronic Test, certification and calibration. Third, it is analyze MRA between Korea and GCC 6countries, SDoC, Mutual Acceptance of International certification such as ILAC(APLAC) and IEC CB scheme.

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 47–52, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80404

This paper intends to draw conclusion and make implication as follows.

First, we must promote FTA and MRA. Second, we make to remove and to easy TBT by MRA between Korea and GCC 6countries. Although the MRA is a system where all parties that have concluded agreement enjoy the advantage, Korea is under a state of concluding only the stage 1 agreements(exchange test documents) with GCC 6 countries. Also, we must conclude MRA stage 2agreements (exchange certification documents)with GCC 6 countries.

2. PREVIOUS STUDIES

2.1 STRATEGY TO REMOVE & EASE TBT IN OECD

SDoC has strengths in cost reduction, time saving and product information protection aspects compared to the certification system while having vulnerability in terms of product safety issue, etc. Therefore, an effective post market surveillance of the regulation authorities must be supported to be operated effectively. WTO's TBT Committee has suggested that the SDoC is more effective TBT elimination method than the MRA (OECD, 2000).

2.2 STRATEGY TO REMOVE & EASE TBT IN APEC TEL MRA

Testing and certification are expensive procedures for exporters, importers and regulators that increase the cost to users and delays the availability of products in a large number of markets.

All stakeholders benefit from simplified procedures that can reduce these costs. At the same time, regulators need to have confidence in the quality of testing that provides the basis for certification of equipment.

In June 1998, the APEC1 Telecommunications and Information Ministers agreed to streamline APEC-wide processes for the testing and type-approval of telecommunications equipment.

This landmark arrangement, the Mutual Recognition Arrangement for Conformity Assessment of Telecommunications Equipment (APEC TEL MRA2), was the first multilateral agreement of its type in the world.

This Arrangement streamlines the Conformity Assessment Procedures for a wide range of telecommunications and telecommunications-related equipment and facilitates trade among the APEC member economies.

It reduces a significant barrier to what is projected to be a US$60 billion industry by 2010.Its scope includes all equipment subject to telecommunication regulations, including wire line and wireless, terrestrial and satellite equipment. For such equipment, the MRA covers electromagnetic compatibility (EMC), specific absorption rate (SAR) and electrical safety aspects as well as purely telecommunications aspects of the conformity assessment requirements.

3. STRATEGY TO REMOVE/EASE OF THE TBT

TBT is an abbreviation for‘ Technical Barriers to Trade’ while this stands for the various obstacles in terms of trade that hinder the free movement of goods and services as the trading

partner countries adopt and apply different Technical Regulations, Standards, Test &Certification Procedures and Inspection Systems, etc.

3.1 STRATEGY TO REMOVE/EASE OF THE TBT

TBT makes the countries to harmonize technical regulations, standards or conformity assessments with the international standards and does not occur in case of being transparent. However, the fact is that TBT occurs if a specific country does not comply with the principles above during legislation and amendment of the laws related to technical regulations, standards or conformity assessments while STC must be submitted to settle this TBT. In the conformity assessment of ICT section, various methods of solution exist on TBT depending on the issue other than filing a lawsuit to WTO if a specific country operates the conformity assessment section differently from TBT.

(1) Request for Introduction of SDoC System

SDoC system stands for the one to guarantee market autonomy and raise efficiency of restriction as a system for the supplier to guarantee by evaluating whether its own product is appropriate for the concerned standard by escaping from the compulsory certification system which requires certification in relation to the product manufacture. Since SDoC(Supplier's Declaration of Conformity) is a follow-up and legal system, it is the method of releasing new products under the manufacturer's own responsibility to become responsible for various problems to follow.

(2) Strategy of MRA

The manufacturers of industrial products are able to export only after acquiring a compulsory standard certification mark. While MRA is concluded in order to save cost and time required for this, only the test report implemented at the exporting country is recognized if MRA stage 1(exchange test documents) is concluded while both the test report and the certification market may be implemented at the exporting country may be implemented if MRA stage 2(exchange certification documents) is concluded.

If both countries conclude the MRA such as FTA, it would be opening the homeland market to the manufacturer of partner country since it is customs-free.

3.2 DOMESTIC ELECTRIC & ELECTRONIC CERTIFICATION SYSTEM AND RELATED LAWS

▪ Conformity Assessment System of Korea

Supplier's Declaration of Conformity (SDoC) the one to guarantee market autonomy and raise efficiency of restriction as a system for the supplier to guarantee by evaluating whether its own product is appropriate for the concerned standard by escaping from the conventional compulsory certification system which requires certification in relation to the product manufacture**.**

(1) Acceptance of Internationally Certified Test Report

In addition to the method of concluding an MRA, various methods to recognized the test reports estimated at the partner country or a third country exist. Among them, the most widely used method is the one to accept test reports of the testing agencies that have been recognized by ILAC(APLAC) and CB Scheme. Test & Certification Based Infrastructure Setup Support

3.3 COMPARATIVE ANALYSIS AND STRATEGY OF REMOVE AND EASE TBT

The systems mentioned above have different characteristics from each other. If the comparative analysis is performed from the perspectives of scope of effect, intensity of effect and usage status in Korea, they can be summarized as follows.

[Figure 1] Comparative Analysis and strategy of the remove and ease TBT

	MRA	**SDoC**	**Mutual Acceptance of International Certification**		**Infrastructure Setup Support Project**
			ILAC(APLAC)	**CB Scheme**	
Scope of Effect	Partner country of agreement	All countries	Participating countries	Participating countries	Beneficiary countries
Intensity of Effect	In stages	Limited to the products that have applied the system	By accepted field	By accepted field	Different according to the supported standard
Current Status of Usage In Korea	Completed stage 1 conclusion with 5 countries and negotiating with a number of countries	Applied to the products with low level of harm	Field of private sector standards	EMC field is not used	Under support
Remarks	Need to promote upper stage with more countries	Necessity to extend applied products is low in a short run	Handle flexibly depending on the acceptance situation of foreign countries	Handle flexibly depending on the acceptance situation of foreign countries	Need to extend support

If both countries conclude the MRA such as FTA, it would be opening the homeland market to the manufacturer of partner country since it is customs-free.

Among APEC members, time and cost required for preparing the copy of agreement can be saved if the MRA Guide prepared by this organization is used. Although Korea has concluded MRA stage 1 with the United States, Canada, Chile and Vietnam, etc., the effect is clearly shown only in the MRA with the United States.

4. CONCLUSION

This study intends to draw conclusions and make policy implications as follows.

First, we must promote a multi-track simultaneous agreements with the countries that have necessity of short-term promotion.

Second, the countries with necessity of short-term promotion on the preferential basis are China, Japan and USA, etc.

Third, it is necessary to conclude MRA agreement with the leading countries among the GCC 6 countries on the preferential basis. It is necessary to prepare negotiation on the preferential basis with GCC 6 countries.

Fourth, support on the countries that have not fully prepared the conformity assessment system needs to be gradually extended. However, the method of support on these countries also must vary depending on the country. KTC (Korea Testing Certification) written MRA GCC test & certification company in 2015 years. So Korea exporting company can export to GCC goods with attached test & certification documents in Korea. To conclude, it is suggest MRA for the remove and reduce TBT to increase export and import among countries.

REFERENCES

[1] Beghin and Bureau (2001). “Quantification of Sanitary, Phytosanitary, and Technical Barriers to Trade forTrade policyAnalysis”,Working Paper, Center for Agricultural and Rural Development, Iowa State University. pp102-119

[2] Johnson, C. (2008). Technical Barriers to Trade: Reducing the Impact of Conformity Assessment Measures.USITC Working Paper.19.pp34-37[1]

[3] OECD (1999). An Assessment Of The Costs For International Trade In Meeting Regulatory Requirements pp66-76

[4] OECD (2000). An Assessment Of The Costs For International Trade In Meeting Regulatory Requirements(TD/TC/WP(99)8/FINAL).pp27-31

[5] OECD (2013), Annual Report on the OECD Guidelines for Multinational Enterprises (2013), OECD publishing pp166-177

[6] OECD(2011), OECD Guidelines for Multinational Enterprises 2011 Edition, OECD publishing.pp127-136

[7] P. S. Huh, Y. J. Park and K. S. Lim. Grouping and Priority Setting for the Expected IT Equipment MRA Conclusion Countries through Index Analysis. (2007)「International Regional Research」. 11(1): pp541-561.

[8] UN Global Compact(2014), United Nations Global Compact Strategy 2014, UN Global Compact. pp76-89

[9] Y.J. Chang and J.M. Seo The Impact of Technical Barriers to Trade (TBT) on Bilateral Trade: A Case of Korea.(2014)「InternationalTrade Research」. 19(1): pp10-33.

[10] Y.K. Lee. Empirical Analysis on the Economic Effect of Mutual Recognition Agreement (MRA) among CountriesIn the Data Communication Sector: Focused on the Case of Korea-U.S. Areement. (2014)「National Policy Research28(3):pp1-26.

AUTHORS

ISO TC 68 Member

ISO TC 195 Member

ISO/IEC SC32 WG1 Vice convenor

PERFORMANCE ANALYSIS OF SYMMETRIC KEY CIPHERS IN LINEAR AND GRID BASED SENSOR NETWORKS

Kaushal Shah and Devesh C. Jinwala

Department of Computer Engineering,
S.V.National Institute of Technology, Surat, India.

ABSTRACT

The linear and grid based Wireless Sensor Networks (WSN) are formed by applications where objects being monitored are either placed in linear or grid based form. E.g. monitoring oil, water or gas pipelines; perimeter surveillance; monitoring traffic level of city streets, goods warehouse monitoring. The security of data is a critical issue for all such applications and as the devices used for the monitoring purpose have several resource constraints (bandwidth, storage capacity, battery life); it is significant to have a lightweight security solution. Therefore, we consider symmetric key based solutions proposed in the literature as asymmetric based solutions require more computation, energy and storage of keys. We analyse the symmetric ciphers with respect to the performance parameters: RAM, ROM consumption and number of CPU cycles. We perform this simulation analysis in Contiki Cooja by considering an example scenario on two different motes namely: Sky and Z1. The aim of this analysis is to come up with the best suited symmetric key based cipher for the linear and grid based WSN.

KEYWORDS

Linear and Grid Based Wireless Sensor Networks, Symmetric Key Based Ciphers, Performance Analysis, Contiki Cooja.

1. INTRODUCTION

The Wireless Sensor Networks (WSNs) are considered to be deployed in random or tree based fashion. However, there are applications of WSN that form specific topology like linear or grid based WSN. The applications where objects being monitored are distributed in either linear or square grid inherently form linear and grid based WSN. Examples of the same are:

i. Monitoring the traffic level of city streets.

ii. Monitoring pipelines carrying oil, water or gas.

iii. Monitoring goods in a warehouse.

iv. Perimeter surveillance.

The typical examples of nodes forming a linear network and square grid are as shown in Fig. 1 and 2 respectively. As shown in Fig. 1, there are 7 nodes deployed in a linear fashion and each node has a communication range of 2. Therefore, such network is known as (7, 2) linear network.

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 53–68, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80405

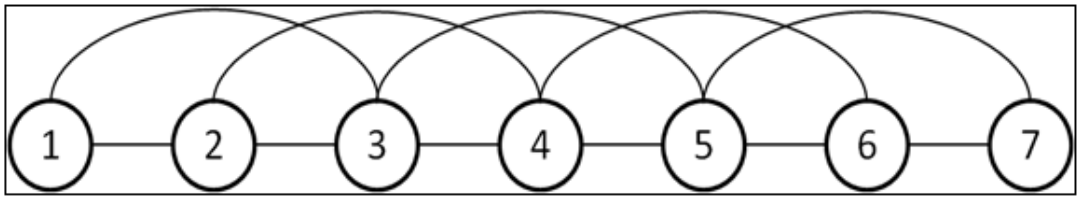


Fig. 1. A (7, 2)-linear sensor network.

As shown in Fig. 2, there are 50 houses of a colony deployed in a square grid manner. Here, each house is considered to have a device that is used for monitoring the energy consumption of the house. These devices pass the aggregated data in a hop by hop manner to the aggregator node, which send the data to the BS. Based on the considered application, there can be multiple aggregator nodes. The same deployment of devices can be considered for the applications of grid based networks. When we consider a single row of houses, it forms a linear network of 10 houses. Therefore, the grid based WSN are formed through the combination of linear networks (when 5 rows are considered, it forms grid based WSN). As the data are passed in hop by hop manner for getting the advantage of aggregation as discussed in [1], the security of the same is a critical issue. The intermediate nodes can alter or passively monitor the data and use it for their own advantage as discussed in [2, 3]. Therefore, the data are required to be encrypted before passing to the next node and the same is discussed in [4-6].

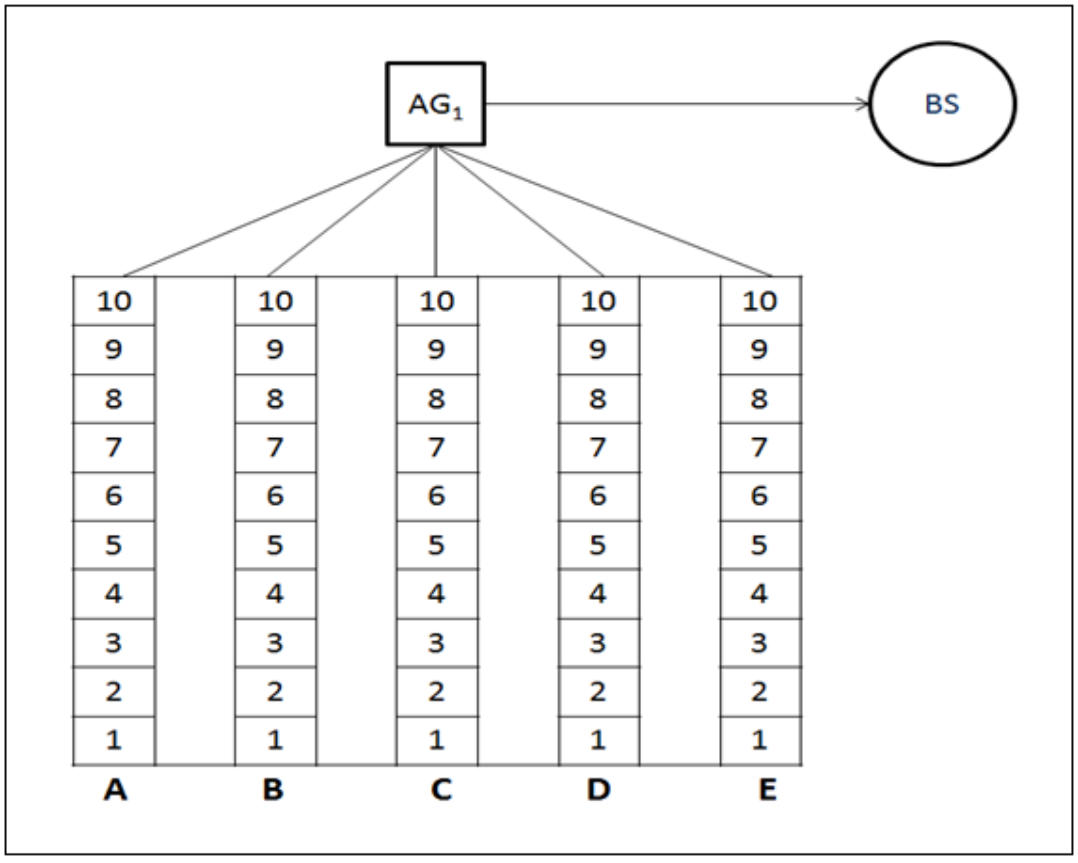


Fig. 2. Considered scenario of colony forming grid based network.

The lightweight schemes for the same are discussed in [7, 8]. However, these schemes do not take the deployment knowledge of sensor nodes under consideration unlike the secure data aggregation scheme for linear WSN proposed in [9]. For protecting the privacy of data, there are schemes proposed in the literature [10-13]. The scheme discussed in [10] is based on the idea of pseudonym changing. The idea of anonymous communication is presented in [11]. The key establishment scheme for the networks monitoring roads is discussed in [14]. The energy efficient communication protocol based on the idea of clustering in road networks is presented in [15]. Using the public key cryptography without using the certificates, the authentication scheme is discussed in [16]. In this paper, we perform the simulation analysis to find out the number of CPU cycles and memory consumption required in providing data security for linear and grid WSN.

The devices (sensor nodes) used in the considered applications are constrained with regard to the computation capabilities, battery power, storage, bandwidth [17, 18]. Therefore, we perform an analysis on the symmetric key based ciphers (as asymmetric key based ciphers require more storage and computation) and analyse the same for the considered example scenario. The performance parameters used in the simulation analysis are: RAM, ROM consumption and number of CPU cycles. The keys are pre-distributed as proposed in [19]. Such analysis is already discussed in the literature [20-22]. However, they have not considered the knowledge of deployment of nodes and the specific topology formed by the application. The main objective of our work is to come up with the best suited cipher for the considered scenario of linear and grid based WSN. Moreover, we perform our analysis on two different motes: sky and z1. The z1 motes have more capability regarding the RAM and the ash memory and the analysis results shows that they require lesser number of CPU cycles as compared to sky motes. In addition, the comparison of two operating systems contiki cooja and tinyOS for sensor nodes is discussed in the paper. We perform a detailed analysis for one of the most significant cryptographic primitives of WSNs: Symmetric Key Block Cipher. We consider the performance parameters, storage and energy for a set of candidate lightweight ciphers. Analysing the performance of the symmetric key based ciphers with respect to the performance parameters: number of CPU cycles and RAM,ROM in contiki cooja, is the major contribution of this paper.

1.1 Organization of the Paper

The rest of the paper is organized as follows: In Section 2, brief introduction about the examined symmetric key based ciphers is discussed. In Section 3, we look at the simulation setup and the methodology for the evaluation of the ciphers. Moreover, the comparison of the operating systems contiki cooja and tinyOS is discussed in this section. In Section 4, we show the simulation results and discuss the same. We summarize our work with conclusions in Section 5.

2. THE SYMMETRIC KEY BASED CIPHERS: EXAMINED

In this section, we discuss block ciphers that are lightweight in nature. Different ciphers are based on different structures like Substitution Permutation Network (SPN), Feistel or Lai-Massey. AES [23]; KLEIN [24]; LED [25]; PRESENT [26] are based on SPN structure. HIGHT [27]; LBlock [28]; MIBS [29]; PICCOLO [30]; SEA [31]; SIMON [32]; TWINE [33] are based on Feistel structure. SPECK [32] is based on ARX (Add-Rotate-Xor) structure. IDEA [34] is based on Lai-Massey structure. We select these ciphers for the analysis as the applications we considered require lightweight solution and with the analysis of the paper we come up with the best suited cipher from the considered lightweight ciphers. We provide an overview of the examined ciphers and the attacks that are possible on each of them. We do not go in the designing details of the ciphers as our main focus is to analyse the performance in terms of CPU cycles and RAM, ROM consumption.

2.1. Substitution Permutation Network (SPN) Structure and Related Ciphers

SPN structure [35] takes a plaintext block and a key as inputs, and performs exchanging "rounds" of substitution (S) and permutation boxes (P-boxes) respectively to deliver the ciphertext block. An S-box substitutes a block of bits given as input by another block of bits as the output. This substitution must guarantee invertibility. A P-box is a permutation of all the bits: it takes input from the outputs of all the S-boxes of previous round, applies permutation of bits, and augments them into the S-boxes of the following round. The key is combined using some group operation like XOR at each round. The S-boxes and P-boxes transformations are efficient to perform in device (like sensors), E.g. exclusive or (XOR) and bitwise rotation.

The following ciphers are based on SPN structure:

Advanced Encryption Standard (AES). We analyse two different implementations of Advanced Encryption Standard (AES) ciphers. One is publicly available and other is designed by contiki cooja developers. Contiki has LLSec (Link Layer Security) layer. This layer is hardware independent, as it uses generic AES driver API instead of directly accessing the hardware. There are multiple AES drivers implemented in Contiki - software-only version and a couple of hardware accelerated ones, including for CC2420 (the radio chip on Sky mote). Authors of [36] show a possible attack on AES, known as biclique cryptanalysis. It uses the concept of exhaustive search on the key with an improvement by linking the keys through key schedule. This attack takes a time complexity of $2^{126.2}$ AES encryptions on the data amount 2^{88}. The other possible attack is meet-in-the-middle [37] that takes less than 2^{100} data/time/memory complexity.

KLEIN. We analyse two different implementations of this cipher: KLEIN64 and KLEIN96. Both the implementations take 64 bits block size. Key lengths are 80 and 96 bits respectively. The number of rounds can either be 12, 16, or 20. The possible attack on this cipher is chosen plaintext key recovery as discussed in [38].

LED. We analyse two different implementations of this cipher: LED64 and LED128. Both the implementations take 64 bits block size. Key lengths are 64 and 128 bits respectively. The number of rounds is 32 and 48 respectively. This cipher does not use key schedule and this is the main difference from other ciphers. The XORing of key is done after every four rounds instead of key schedule. The number of rounds of this cipher is more as compared to other ciphers for compensating the key schedule. The differential cryptanalysis results on this cipher are discussed in [39]. The attacks on LED64 can be reduced to 12 and 16 rounds is described by the authors. The other possible attack is meet-in-the-middle as discussed in [40]. The complexity of the attack on 8 rounds of LED64 and 16 rounds of LED128 is lesser as compared to exhaustive key search.

PRESENT. It is the most popular cipher among all lightweight block ciphers. We analyse two different implementations of this cipher: PRESENT Size and PRESENT Speed. Both implementations take 64 bits block size. Key lengths are either 80 or 128 bits with number of rounds as 31. There many cryptanalysis results as discussed in [41-43]. Authors in [44] discuss about two bicliques possible on two implementation of PRESENT.

2.2. Feistel Structure and Related Ciphers

Feistel structure [45] takes plaintext block as input and divides it into two halves, L (left) and R (right). R half is given as input to a feistel function along with the round key. It is also used as an L half for the next round. Output of the feistel function is XORed with L half and used as an R half for the next round. The same process is repeated till last round. The advantage with this structure is, just by reversing the key schedule decryption can be done.

The following ciphers are based on feistel structure (or a modified feistel structure):

HIGHT. This cipher uses block size of 64 and key length of 128 bits. It uses 32 rounds and sometimes uses modular addition instead of XOR operation. The biclique attack against HIGHT is proposed in [46]. Moreover, the differential cryptanalysis attack is described in [47].

LBLOCK. This cipher uses block size of 64 and key length of 80 bits. The number of rounds is 32 and the usage of 8 S-boxes and permutation of 4 bits are applied. The biclique attack against LBLOCK is proposed in [48]. The authors also discuss the prevention of this attack with the help of modified key schedule algorithm.

MIBS. We analyse two different implementations of this cipher: MIBS64 and MIBS80. Both implementations take 64 bits block size. Key lengths are 64 and 80 bits respectively. The linear attacks on MIBS are discussed in [49]. The authors show the differential cryptanalysis on 14 rounds, ciphertext only attacks on 13 rounds and an impossible differential attack on 12 rounds of MIBS.

PICCOLO. We analyse two different implementations of this cipher: PICCOLO80 and PICCOLO128. Both implementations take 64 bits block size. Key lengths are 80 and 128 bits respectively. It uses two feistel functions. It requires less than 1000 gates when implemented on hardware. Authors of [44] discuss the biclique attacks on both the implementations of PICCOLO.

SEA. This cipher uses n bits block size. The value of n can be 48, 96, or 144 bits. It uses a two branch feistel structure as modified feistel structure. The security analysis is discussed in [31].

SIMON. This cipher uses different block sizes like 32, 48, 64, 96, 128 bits. It is based on a balanced feistel network. The key size can be 64, 72, 96, 128, 144, 192, 256 bits. It is optimized for the hardware implementations. The differential cryptanalysis is possible on this cipher as discussed in [50, 51].

TWINE. We analyse two different implementations of this cipher: TWINE80 and TWINE128. Both the implementations take 64 bits block size. Key lengths are 80 and 128 bits respectively. The number of rounds is 36 in both the implementations. The feistel function uses a single Sbox and subkey addition. This function is repeated 8 times in each round. Two biclique attacks on two implementations are discussed in [52].

2.3. Add-Rotate-Xor Structure and Related Cipher

This structure involves 3 operations:

1. Modular Addition

2. Rotation with fixed rotation amounts

3. XOR

These ARX operations are immune to timing attacks because they run in defined constant time. As these operations are fast and cheap in hardware and software, the ciphers based on ARX operations are popular.

SPECK. This cipher uses different block sizes like 32, 48, 64, 96, 128 bits. It is based on an Add-Rotate-Xor (ARX) structure. The key size can be 64, 72, 96, 128, 144, 192, 256 bits. It is

optimized for the software implementations. The differential cryptanalysis is possible on this cipher as discussed in [50, 51].

2.4. Lai-Massey Structure and Related Cipher

As shown in Fig. 3, Lai-Massey Structure [53] divides the plaintext in two equal halves L_0 and R_0 as input. Two round functions are used; H and F. Keys are used with function F. The output of function H is given as input to function F.

$$(L_{i+1}{}', R_{i+1}{}') = H(L_i{}' + T_i, R_i{}' + T_i)$$
$$where\, T_i = F(L_i{}' - R_i{}', K_i)\, and\, (L_0{}', R_0{}') = H(L_0, R_0)$$

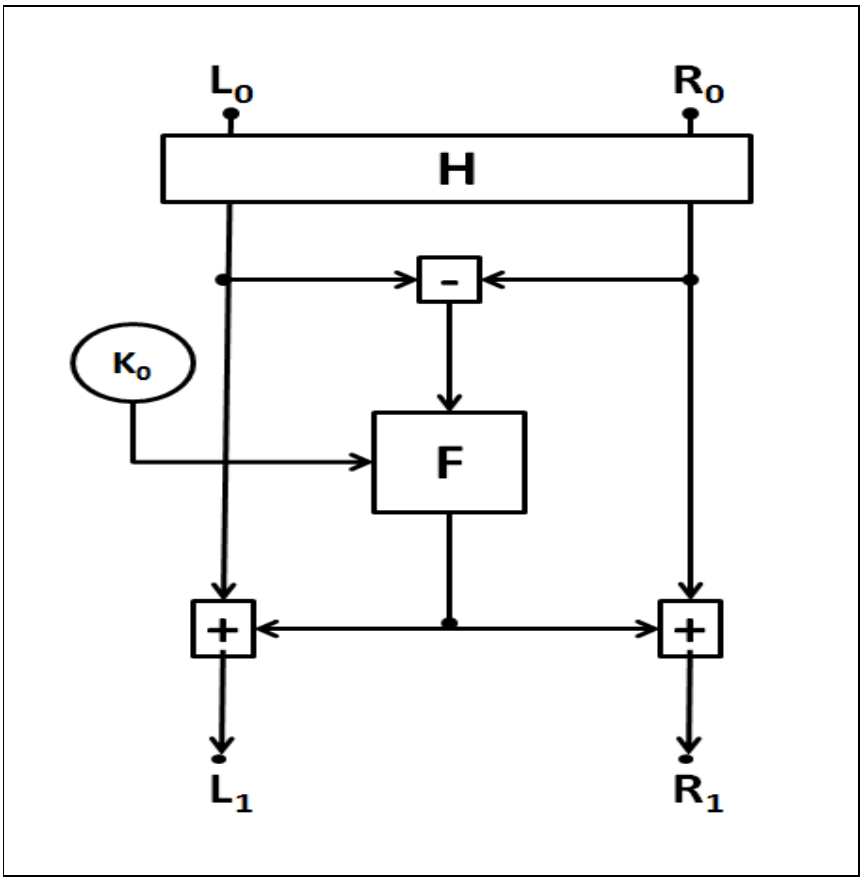


Fig. 3. Lai-Massey Structure

IDEA. This cipher uses 64 bits block size and 128 bits key. It is composed of 8.5 rounds. A final round that is "half round", comes after eight rounds and used for the output transformation (the swap of the centre two values counterbalances the swap toward the end of the last round, so that there is no net swap). It is included in the package PGP (Pretty Good Privacy). There is six rounds attack that exploits key schedule of IDEA with linear cryptanalysis as discussed in [54]. The biclique framework is used by the authors of [55] to speed up the key recovery.

The summary of all the examined symmetric key based block ciphers with regard to the structure they are built on, block size, key size, and the attacks that are possible on each of them is shown in Table 1. From all the examined ciphers, SIMON and SPECK are considered to be the block ciphers for IoT (Internet of Things) environment as discussed in [56].

Table 1. Summary of Block Ciphers Examined.

Sr. No.	Cipher	Refer-ence	Block Size (bits)	Key Size (bits)	Structure	Possible Attacks
1	AES	[23]	128	128	SPN	- Biclique cryptanalysis [36] - Meet-in-the-middle [37]
2	HIGHT	[27]	64	128	Fiestel	- Differential cryptanalysis [47]
3	IDEA	[34]	64	128	Lai-Massey	- Linear cryptanalysis [54] - Biclique [55]
4	KLEIN	[24]	64	64,96	SPN	- Chosen plaintext [38]
5	LBLOCK	[28]	64	80	Fiestel	- Biclique [48]
6	LED	[25]	64	64,128	SPN	- Differential cryptanalysis [39] - Meet-in-the-middle [40]
7	MIBS	[29]	64	64,80	Fiestel	- Differential cryptanalysis [49]
8	PRESENT	[26]	64	80	SPN	- Cryptanalysis [41-43] - Biclique [44]
9	PICCOLO	[30]	64	80,128	Fiestel	- Biclique [44]
10	SEA	[31]	96	96	Fiestel	- Cryptanalysis [31]
11	SIMON	[32]	32,48,64, 96,128	64,72,96,128, 144,192,256	Fiestel	- Differential Cryptanalysis [50, 51]
12	SPECK	[32]	32,48,64, 96,128	64,72,96,128, 144,192,256	ARX	- Differential Cryptanalysis [50, 51]
13	TWINE	[33]	64	80,128	Fiestel	- Biclique [52]

3. SIMULATION SETUP AND THE METHODOLOGY FOR EVALUATION

We analyse the number of CPU cycles and RAM, ROM consumption in achieving data security for the secure data aggregation scheme proposed in [9]. The scheme uses the pre distributed keys (proposed in [19]) for the purpose of encryption. We work with the nodes that are constrained regarding:

- Storage
- Communication range (It is assumed that nodes can communicate at least till one hop)
- Battery life

Contiki Cooja is a simulator specifically designed for IoT devices that are having the constraints as described. It is also used as an emulator because the code to be executed by the node is the exact same firmware one may upload to physical nodes [57].

3.1. Simulation Setup

This section discusses the simulation results for the considered example scenario. We focus on the data security as it is crucial when designing a data aggregation scheme. Passively acquired data can be used for malicious purpose if confidentiality of data is not taken care of. Our criteria are to measure the number of CPU cycles and RAM, ROM consumption for providing data security. In contiki cooja, there are several options regarding the selection of devices for which one wants to emulate. E.g. Cooja mote, MicaZ mote, CC430 mote, Z1 mote, Sky mote, etc. The comparison between the operating systems TinyOS and Contiki is discussed in [58]. In TinyOS, the application has to be replaced totally when the code is changed. However, the contiki OS is better when it comes to updating the deployed application as it can dynamically replace the changed programs. The protocols discussed in [9, 59-60] require the code to be updated every time the value of N (total number of nodes in the network) or k (number of consecutive nodes) is changed. Moreover, Contiki supports dynamic loading and unloading of the code and multi-threading. Contiki is an event driven OS and event handlers cannot pre-empt each other. However, interrupts can pre-empt the current running process.

We use Sky motes and Z1 motes of contiki cooja for the purpose of simulation. Sky mote features a 16-bit MSP430 MCU, 10 kB RAM, 48 kB ROM, a cc2420 802.15.4 radio transceiver, an external Flash memory, and temperature, humidity and brightness sensor [17]. Z1 mote has higher configuration and uses MSP430F2617 MCU [18]. The specifications considered for Sky and Z1 motes are as shown in Table 2 and Table 3 respectively.

Table 2. Sky mote Specifications

Flash Memory	48 KB
RAM	10 KB
Current Consumption	20 mA
Operating Voltage	3 V
Micro-controller	MSP430

Table 3. Z1 mote Specifications

Flash Memory	92 KB
RAM	8 KB
Current Consumption	19.7 mA
Operating Voltage	3 V
Micro-controller	MSP430F2617

3.1. Methodology for Evaluation of Ciphers

In this section, we analyse the symmetric key based block ciphers with regard to the number of CPU cycles, energy and memory they consume if applied on Sky and Z1 Motes. The energy is calculated through following steps:

- Add the header file #include “energest.h" in a .c file of the considered cipher.

- To get CPU cycles involved in the different ciphers, add “printf(“energy cpu: %lu", energest_type_time(ENERGEST_TYPE_CPU));" line in PROCESS_THREAD of the .c file.

- To get the power consumption, the formula is: Power(mW) = $\frac{r_x ON}{CPU + LPM} * 20mA * 3V$

Sky and Z1 motes have 20mA current value and 3V voltage. Therefore, CPU cycles received from running the code on Sky or Z1 motes, if multiplied with 60 will give the power consumption in Watts. When this value is multiplied with simulation time, it gives energy consumption in joules. For getting the number of CPU cycles involved in ciphers, we run each cipher separately and follow the steps as discussed. Fig. 4 shows the simulation result, when we run a TWINE cipher on a grid of 20 Sky motes.

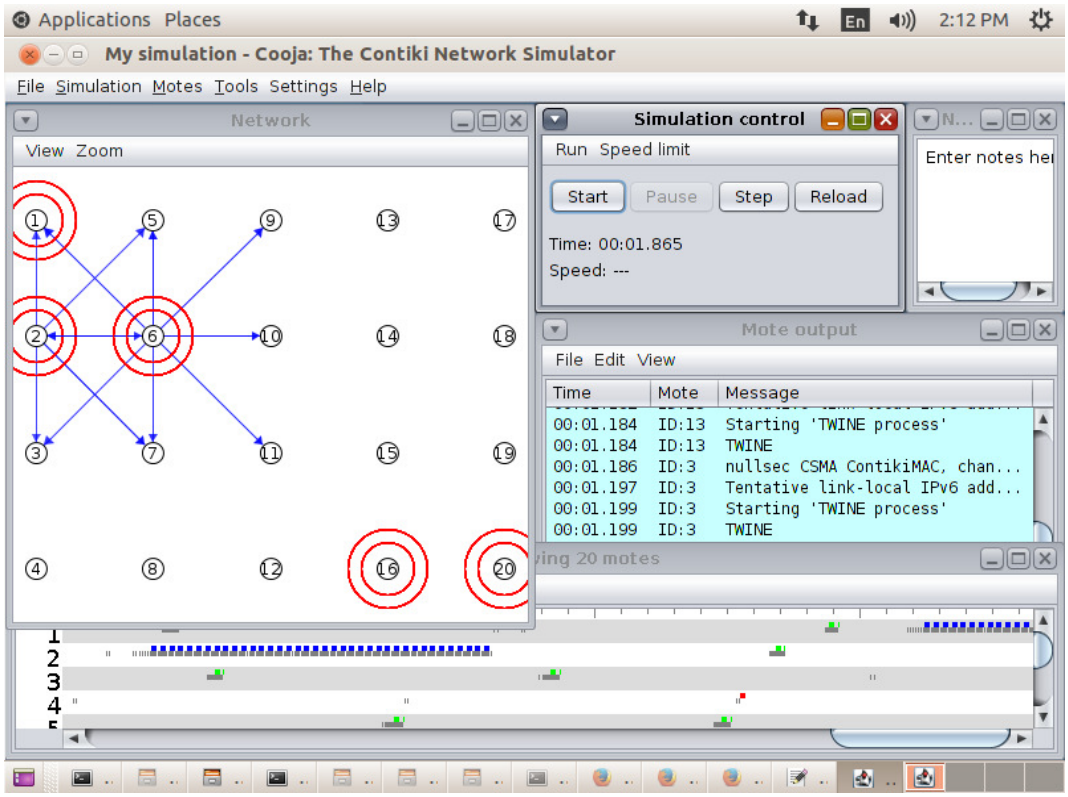


Fig. 4. Simulation of Twine80 Cipher with Sky Mote

In order to obtain the memory (RAM, ROM) consumption, we use “size" command. Fig. 5 shows the use of size command. Here, .text column refers to the ROM consumption by different ciphers in bytes. The .data and .BSS columns show the RAM consumption. We run all the ciphers in the same manner on both Sky and Z1 motes respectively and the results are as shown in Tables 4 and 5.

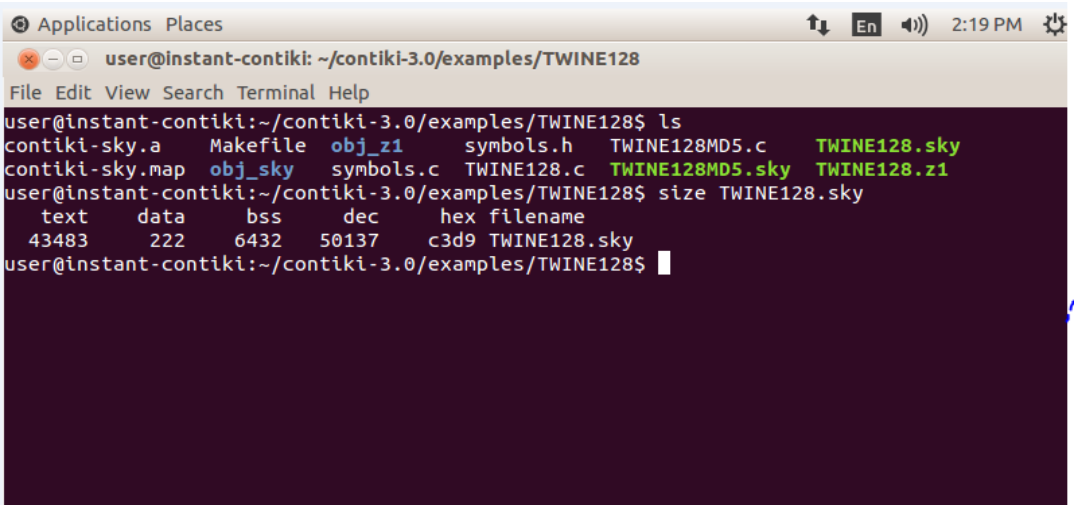


Fig. 5. Size command on Twine Cipher with Sky Mote

4. SIMULATION RESULTS

We can see from the Table 4 that, each cipher requires different amount of CPU cycles and RAM, ROM consumption in the grid network of 20 sky motes. The one that requires maximum number of CPU cycles is LED128 (11117) and the one that requires minimum number of CPU cycles is KLEIN64 (1401). The AES designed by contiki developers specifically for sky motes uses hardware acceleration that helps in reducing number of CPU cycles compared to publicly defined AES (from Table 4, we can see AES (Contiki) requires 1503 whereas AES (Public) requires 1582 number of CPU cycles). SPECK (128 bits block and key size) cipher is designed specifically for resource constrained environments requires 1403 number of CPU cycles.

Table 4. Sky Mote: No. of CPU Cycles and RAM, ROM

Cipher	CPU Cycles	RAM,ROM (bytes)
AES (Contiki)	1503	49973
AES (Public)	1582	51329
HIGHT	1461	50093
IDEA	2375	50151
KLEIN64	1401	50719
KLEIN96	1562	50763
LBLOCK	1404	50555
LED64	7819	50149
LED128	11117	50133
MIBS64	1559	50381
MIBS80	1632	50947
PRESENT_Size	4224	51187
PRESENT_Speed	3715	51251
PICCOLO80	1487	50055
PICCOLO128	1512	50111
SEA	1665	49923
SIMON128	1808	50731
SPECK128	1403	49995
TWINE80	1668	49943
TWINE128	1716	50137

Table 5. Z1 Mote: No. of CPU Cycles and RAM, ROM

Cipher	CPU Cycles	RAM,ROM (bytes)
AES (Contiki)	633	49279
AES (Public)	285	49131
HIGHT	167	48409
IDEA	643	48453
KLEIN64	184	48995
KLEIN96	262	49043
LBLOCK	176	48479
LED64	2899	48499
LED128	4342	48483
MIBS64	242	48421
MIBS80	277	48631
PRESENT_Size	1300	49021
PRESENT_Speed	896	49043
PICCOLO80	237	48383
PICCOLO128	251	48451
SEA	310	48319
SIMON128	373	49079
SPECK128	128	48347
TWINE80	307	48333
TWINE128	336	48531

When we run all the ciphers on the Z1 mote, it takes a lesser number of CPU cycles as we can see from Table 5 (AES (pub) on sky mote takes 1582 CPU cycles, whereas on Z1 mote it takes 285 CPU cycles). The AES code provided by contiki cooja developers is specifically designed for Sky mote by using hardware acceleration. Therefore, the RAM consumption of the same is lesser as shown in Table 4. The number of CPU cycles for Z1 motes is always lesser compared to Sky motes, as the configuration of Z1 mote is superior concerning the flash memory that can be used as either RAM or ROM. Therefore, the number of CPU cycles for running different ciphers is lesser for Z1 motes compared to Sky motes as shown in Tables 4 and 5.

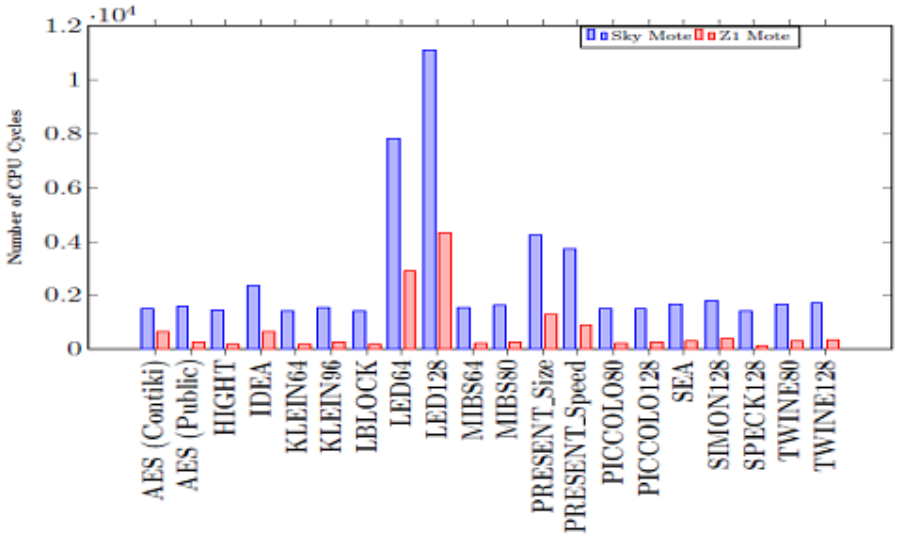


Fig 6. CPU Cycles

The comparison of the CPU cycles required by both the motes is as shown in Fig. 6. It shows that Z1 mote takes lesser number of CPU cycles as compared to Sky mote for all the ciphers. The comparison in terms of RAM and ROM consumption of both the motes is as shown in Fig. 7. It shows that Z1 mote takes lesser amount of RAM, ROM consumption as compared to Sky mote in running all the ciphers.

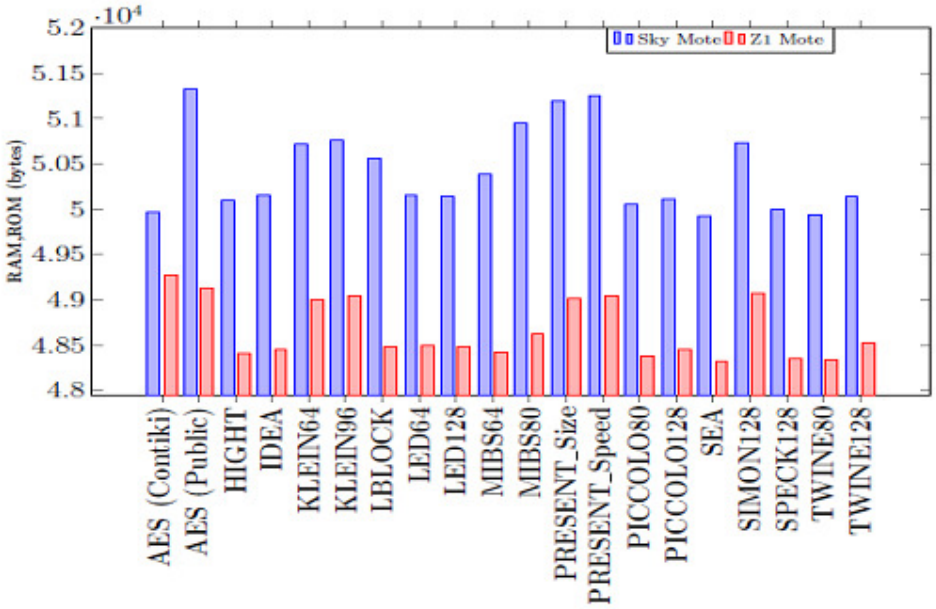


Fig 7. RAM, ROM

5. CONCLUSIONS

We analysed symmetric key based block ciphers on two different motes (sky and z1) in contiki cooja. This analysis concerning the number of CPU cycles and RAM, ROM consumption helps in deciding which cipher can be used for the secure data aggregation scheme in different scenarios. In the constrained scenario of sensor nodes, it is better to use lightweight ciphers such as HIGHT; KLEIN; PICCOLO; SIMON; SPECK or TWINE. The hardware accelerated AES cipher by contiki cooja uses minimum number of CPU cycles when we take Sky mote under consideration. Since speed is correlated with energy consumption, SPECK 128/128 is a better choice in energy critical applications as it produces energy efficient solution with an encryption cost of 1403 cycles on Sky mote, or 128 cycles on Z1 mote. We have given a detailed analysis for one of the most significant cryptographic primitives for WSNs: Symmetric Key Block Cipher, by considering the performance parameters, storage and energy for a set of candidate lightweight ciphers. We are working on optimizing the AES cipher regarding speed (reducing the number of CPU cycles) and size (reducing the RAM, ROM consumption). This optimized version of AES will be hardware independent. i.e. it will not depend on the mote under consideration (E.g. sky or z1) and produce the optimal results.

REFERENCES

[1] L. Krishnamachari, D. Estrin, and S. Wicker, “The impact of data aggregation in wireless sensor networks," in Proceedings. 22nd International Conference on Distributed Computing Systems Workshops, IEEE, 2002, pp. 575-578.

[2] S. Ozdemir and Y. Xiao, “Secure data aggregation in wireless sensor networks: A comprehensive overview," Computer Networks, vol. 53, no. 12, pp. 2022-2037, 2009.

[3] A. Perrig, J. Stankovic, and D. Wagner, “Security in wireless sensor networks," Communications of the ACM, vol. 47, no. 6, pp. 53-57, 2004.

[4] W. Du, J. Deng, Y. S. Han, P. K. Varshney, J. Katz, and A. Khalili, “A pairwise key predistribution scheme for wireless sensor networks," ACM Transactions on Information and System Security (TISSEC), vol. 8, no. 2, pp. 228-258, 2005.

[5] O. Yagan and A. M. Makowski, “Modeling the pairwise key predistribution scheme in the presence of unreliable links," IEEE Transactions on Information Theory, vol. 59, no. 3, pp. 1740-1760, 2013.

[6] F. Yavuz, J. Zhao, O. Yagan, and V. Gligor, “On secure and reliable communications in wireless sensor networks: Towards k-connectivity under a random pairwise key predistribution scheme," in International Symposium on Information Theory (ISIT), IEEE, 2014, pp. 2381-2385.

[7] G. Kalogridis, Z. Fan, and S. Basutkar, “Affordable privacy for home smart meters," in Ninth International Symposium on Parallel and Distributed Processing with Applications Workshops (ISPAW). IEEE, 2011, pp. 77-84.

[8] G. De Meulenaer, F. Gosset, F.-X. Standaert, and O. Pereira, “On the energy cost of communication and cryptography in wireless sensor networks," in International Conference on Wireless and Mobile Computing, Networking and Communications. IEEE, 2008, pp. 580-585.

[9] K. Shah and D. C. Jinwala, “A secure expansive aggregation in wireless sensor networks for linear infrasturcture," in Region 10 Symposium (TENSYMP), IEEE, 2016, pp. 207-212.

[10] I. Memon, L. Chen, Q. A. Arain, H. Memon, and G. Chen, “Pseudonym changing strategy with multiple mix zones for trajectory privacy protection in road networks," International Journal of Communication Systems, vol. 31, no. 1, 2018.

[11] I. Memon, I. Hussain, R. Akhtar, and G. Chen, “Enhanced privacy and authentication: An efficient and secure anonymous communication for location based service using asymmetric cryptography scheme," Wireless Personal Communications, vol. 84, no. 2, pp. 1487-1508, 2015.

[12] Q. A. Arain, D. Zhongliang, I. Memon, S. Arain, F. K. Shaikh, A. Zubedi, M. A. Unar, A. Ashraf, and R. Shaikh, “Privacy preserving dynamic pseudonym-based multiple mix-zones authentication protocol over road networks," Wireless Personal Communications, vol. 95, no. 2, pp. 505-521, 2017.

[13] Q. A. Arain, Z. Deng, I. Memon, A. Zubedi, and F. A. Mangi, “Map services based on multiple mix-zones with location privacy protection over road network," Wireless Personal Communications, vol. 97, no. 2, pp. 2617-2632, 2017.

[14] I. Memon, “A secure and efficient communication scheme with authenticated key establishment protocol for road networks," Wireless Personal Communications, vol. 85, no. 3, pp. 1167-1191, 2015.

[15] Q. A. Arain, M. A. Uqaili, Z. Deng, I. Memon, J. Jiao, M. A. Shaikh, A. Zubedi, A. Ashraf, and U. A. Arain, “Clustering based energy efficient and communication protocol for multiple mix-zones over road networks," Wireless Personal Communications, vol. 95, no. 2, pp. 411-428, 2017.

[16] I. Memon, M. R. Mohammed, R. Akhtar, H. Memon, M. H. Memon, and R. A. Shaikh, “Design and implementation to authentication over a GSM system using certificate-less public key cryptography (CL-PKC)," Wireless personal communications, vol. 79, no. 1, pp. 661-686, 2014.

[17] H. Edu, “Tmotesky: Low power wireless sensor module," 2004. [Online]. Available: http://www.eecs.harvard.edu/~konrad/projects/shimmer/references/tmote-sky-datasheet.pdf

[18] Zolertia, “Z1 features: Quick hardware tour," 2013. [Online]. Available: http://zolertia.sourceforge.net/wiki/index.php/Z1

[19] K. A. Shah and D. C. Jinwala, “Novel approach for pre-distributing keys in WSNs for linear infrastructure," Wireless Personal Communications, vol. 95, no. 4, pp. 3905-3921, 2017.

[20] Y. W. Law, J. Doumen, and P. Hartel, “Survey and benchmark of block ciphers for wireless sensor networks," ACM Transactions on Sensor Networks (TOSN), vol. 2, no. 1, pp. 65-93, 2006.

[21] M. Cazorla, K. Marquet, and M. Minier, “Survey and benchmark of lightweight block ciphers for wireless sensor networks," in International Conference on Security and Cryptography (SECRYPT), IEEE, 2013, pp. 1-6.

[22] T. Eisenbarth, Z. Gong, T. Guneysu, S. Heyse, S. Indesteege, S. Kerckhof, F. Koeune, T. Nad, T. Plos, F. Regazzoni et al., “Compact implementation and performance evaluation of block ciphers in attiny devices," Progress in Cryptology- AFRICACRYPT, 2012, pp. 172-187.

[23] N. F. Pub, “197: Advanced encryption standard (AES)," Federal Information Processing Standards Publication, vol. 197, no. 441, pp. 0311, 2001.

[24] Z. Gong, S. Nikova, and Y. W. Law, KLEIN: a new family of lightweight block ciphers. Springer, 2011.

[25] J. Guo, T. Peyrin, A. Poschmann, and M. Robshaw, “The led block cipher," in Cryptographic Hardware and Embedded Systems-CHES 2011. Springer, pp. 326-341.

[26] A. Bogdanov, L. R. Knudsen, G. Leander, C. Paar, A. Poschmann, M. J. Robshaw, Y. Seurin, and C. Vikkelsoe, PRESENT: An ultra-lightweight block cipher. Springer, 2007.

[27] D. Hong, J. Sung, S. Hong, J. Lim, S. Lee, B.-S. Koo, C. Lee, D. Chang, J. Lee, K. Jeong et al., “Hight: A new block cipher suitable for low-resource device," in Cryptographic Hardware and Embedded Systems-CHES 2006. Springer, pp. 46-59.

[28] W.Wu and L. Zhang, “Lblock: a lightweight block cipher," in Applied Cryptography and Network Security. Springer, 2011, pp. 327-344.

[29] M. Izadi, B. Sadeghiyan, S. S. Sadeghian, and H. A. Khanooki, “Mibs: a new lightweight block cipher," in Cryptology and Network Security. Springer, 2009, pp. 334-348.

[30] K. Shibutani, T. Isobe, H. Hiwatari, A. Mitsuda, T. Akishita, and T. Shirai, “Piccolo: an ultra-lightweight blockcipher," in Cryptographic Hardware and Embedded Systems-CHES 2011. Springer, pp. 342-357.

[31] F.-X. Standaert, G. Piret, N. Gershenfeld, and J.-J. Quisquater, “Sea: A scalable encryption algorithm for small embedded applications," in Smart Card Research and Advanced Applications. Springer, 2006, pp. 222-236.

[32] R. Beaulieu, S. Treatman-Clark, D. Shors, B. Weeks, J. Smith, and L. Wingers, “The simon and speck lightweight block ciphers," in 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), 2015, pp. 1-6.

[33] T. Suzaki, K. Minematsu, S. Morioka, and E. Kobayashi, “TWINE: A lightweight block cipher for multiple platforms," in Selected Areas in Cryptography. Springer, 2012, pp. 339-354.

[34] X. Lai, “On the design and security of block ciphers," Ph.D. dissertation, Diss. Techn. Wiss ETH Zurich, Nr. 9752, Ref.: JL Massey; Korref.: H. Buhlmann, 1992.

[35] J. B. Kam and G. I. Davida, “Structured design of substitution-permutation encryption networks," IEEE Transactions on Computers, vol. 100, no. 10, pp. 747-753, 1979.

[36] A. Bogdanov, D. Khovratovich, and C. Rechberger, “Biclique cryptanalysis of the full AES," in Advances in Cryptology-ASIACRYPT 2011. Springer, pp. 344-371.

[37] P. Derbez, P.-A. Fouque, and J. Jean, “Improved key recovery attacks on reduced round AES in the single-key setting," in Advances in Cryptology-EUROCRYPT 2013. Springer, pp. 371-387.

[38] J.-P. Aumasson, M. Naya-Plasencia, and M.-J. O. Saarinen, “Practical attack on 8 rounds of the lightweight block cipher klein," in Progress in Cryptology-INDOCRYPT 2011. Springer, pp. 134-145.

[39] F. Mendel, V. Rijmen, D. Toz, and K. Var_c_, “Differential analysis of the LED block cipher," in Advances in Cryptology-ASIACRYPT 2012. Springer, 2012, pp. 190-207.

[40] T. Isobe and K. Shibutani, “Security analysis of the lightweight block ciphers XTEA, LED and Piccolo," in Information Security and Privacy. Springer, 2012, pp. 71-86.

[41] J. Nakahara Jr, P. Sepehrdad, B. Zhang, and M.Wang, “Linear (hull) and algebraic cryptanalysis of the block cipher present," in Cryptology and Network Security. Springer, 2009, pp. 58-75.

[42] B. Collard and F.-X. Standaert, “A statistical saturation attack against the block cipher present," in Topics in Cryptology-CT-RSA 2009. Springer, pp. 195-210.

[43] G. Leander, “On linear hulls, statistical saturation attacks, present and a cryptanalysis of puffin," in Advances in Cryptology-EUROCRYPT 2011. Springer, pp. 303-322.

[44] K. Jeong, H. Kang, C. Lee, J. Sung, and S. Hong, “Biclique cryptanalysis of lightweight block ciphers present, piccolo and led." IACR Cryptology ePrint Archive, vol. 2012, p. 621, 2012.

[45] Tutorialspoint, “Fiestel block cipher," 2016. [Online]. Available: http://www.tutorialspoint.com/cryptography/feistel_block_cipher.htm

[46] D. Hong, B. Koo, and D. Kwon, “Biclique attack on the full HIGHT," in Information Security and Cryptology-ICISC 2011. Springer, pp. 365-374.

[47] J. Chen, M. Wang, and B. Preneel, “Impossible differential cryptanalysis of the lightweight block ciphers tea, xtea and hight," in Progress in Cryptology- AFRICACRYPT 2012. Springer, 2012, pp. 117-137.

[48] Y. Wang, W. Wu, X. Yu, and L. Zhang, “Security on Lblock against biclique cryptanalysis," in Information Security Applications. Springer, 2012, pp. 1-14.

[49] A. Bay, J. Nakahara Jr, and S. Vaudenay, “Cryptanalysis of reduced-round MIBS block cipher," in Cryptology and Network Security. Springer, 2010, pp. 1-19.

[50] A. Biryukov, A. Roy, and V. Velichkov, “Differential analysis of block ciphers simon and speck," in International Workshop on Fast Software Encryption. Springer, 2014, pp. 546-570.

[51] F. Abed, E. List, S. Lucks, and J. Wenzel, “Differential cryptanalysis of round reduced simon and speck," in International Workshop on Fast Software Encryption. Springer, 2014, pp. 525-545.

[52] M. Coban, F. Karakoc, and O. Boztas, “Biclique cryptanalysis of twine," in Cryptology and Network Security. Springer, 2012, pp. 43-55.

[53] S. Vaudenay, “On the lai-massey scheme," in Advances in Cryptology- ASIACRYPT99. Springer, 1999, pp. 8-19.

[54] E. Biham, O. Dunkelman, and N. Keller, “A new attack on 6-round idea," in Fast Software Encryption. Springer, 2007, pp. 211-224.

[55] D. Khovratovich, G. Leurent, and C. Rechberger, “Narrow-bicliques: cryptanalysis of full IDEA," in Advances in Cryptology-EUROCRYPT 2012. Springer, 2012, pp. 392-410.

[56] R. Beaulieu, D. Shors, J. Smith, S. Treatman-Clark, B. Weeks, and L. Wingers, “Simon and speck: Block ciphers for the internet of things." IACR Cryptology ePrint Archive, vol. 2015, p. 585, 2015.

[57] F. Osterlind, A. Dunkels, J. Eriksson, N. Finne, and T. Voigt, “Cross-level sensor network simulation with cooja," in Proceeding of 31st IEEE Conference on Local Computer Networks, 2006, pp. 641-648.

[58] T. Reusing, “Comparison of operating systems tinyos and contiki," Sens. Nodes Operation, Netw. Appli.(SN), vol. 7, 2012.

[59] K. Shah and D. C. Jinwala, “Expansive aggregation in wireless sensor networks for linear infrastructure," 3rd Security and Privacy Symposium, IIIT – Delhi, 2015.

[60] K. A. Shah and D. C. Jinwala, “Privacy preserving, verifiable and resilient data aggregation in grid based networks," The Computer Journal, pp. 1-15, 2018. DOI: 10.1093/comjnl/bxy013

AUTHORS

Kaushal Shah is a PhD research scholar in Computer Engineering at the Department of Computer Engineering, S. V. National Institute of Technology, Surat, India. He has received his M.E. degree in Computer Science and Engineering from Government Engineering College, Modasa, India. His research interests broadly include Information Security, Wireless Sensor Networks and Protocol designing.

Devesh Jinwala has been working as a Professor in Computer Engineering at the Department of Computer Engineering, S. V. National Institute of Technology, Surat, India since 1991. His principal research areas of interest are broadly Security, Cryptography, Algorithms and Software Engineering. Specifically his work focuses on Security and Privacy Issues in Resource constrained environments (Wireless Sensor Networks) and Data Mining, Attribute-based Encryption techniques, Requirements Specification, and Ontologies in Software Engineering. He has been/is the Principal Investigator of several sponsored research projects funded by ISRO, GUJCOST, Govt of Gujarat and DiETY-MCIT-Govt of India.

PERTURBED ANONYMIZATION: TWO LEVEL SMART PRIVACY FOR LBS MOBILE USERS

Ruchika Gupta, Udai Pratap Rao and Manish Kumar

Department of Computer Engineering, S. V. National Institute of Technology, Surat, Gujarat, India

ABSTRACT

The use of smart mobile devices like tablets, smart phones and navigational gadgets provide most promising communication and better services to mobile users. Location Based Services (LBS) have become very common in recent years. Mobile users submit their location dependent queries to the untrusted LBS server to acquire a particular service. Ideally, user's personal information such as location data is supposed to be protected while communicating to LBS and at the same time quality of service must be maintained. Therefore, there is a need to have a balanced trade-off between privacy and quality of service. To fulfil such trade-off, this paper proposes a solution that first forms the cloaking region at mobile device, perform perturbation to handle the problem of trusted third party and the anonymizer further anonymizes the location to remove the problem of enough users required to form the cloaking region. The proposed approach protects the location privacy of the user and also maintains the quality of service by selecting appropriate service to the particular user. The proposed algorithm provides two-level location protection to the user, and thus ensures smart mobility of the LBS user.

KEYWORDS

Location Privacy, Location Based Services, Perturbation, Anonymity, Point of Interest, Smart Mobility

1. INTRODUCTION

As mobile commerce is growing at a fast rate, there is a huge demand of location services by the mobile users. As a result, the user's personal information is vulnerable to the privacy breach. Almost all location services demand exact location of the user to provide accurate services in return. Location based services (LBS) continue to grow with the maturity of the positioning technology like global positioning system (GPS) [1]. LBS can be initiated when a predefined event occurs, for e.g. occurrence of an event when a user approaches or leaves the point of interest (POI). Due to the frequent exchange of private information, effective mechanisms are needed for positioning management and protection of the location data. It is agreed that users of LBS demand to have complete control over their location data due to its high vulnerability to location privacy attacks [2]. Ideally, LBS provider must provide the features so that users can manage their location information and can decide with whom and under what conditions their private information can be disclosed. The user shares her location with location server to gain personalized services in return. These services include the discovery of POIs, beforehand traffic information, route assistance and the like. The shared private information can be misused if heard by an adversary. For instance, a user requests a list of highly specialized medical care providers informing about her medical condition to the service provider who may indirectly reveals user's

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 69–79, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80406

medical condition and may misuse them later for some personal gains. The server processes the user's query and returns a set of POIs back. Sometimes, there can be the case when an LBS server promotes a business and in return of few candidate results, it also offers coupons to the user against multiple queries sent as shown in Fig 1. The main drawback of such approach is that the server returns too much unnecessary POIs that leak our important information of the database.

Fig 1: Query Processing Example

Trusted Third Party (TTP) based LBS solution has a major disadvantage of single point of failure and can also be viewed as a single promising point of attack by an adversary. Privacy is at stake if an adversary somehow takes control of the third party. On the other hand, mobile based mechanisms face the problem of enough users to form cloaking region and introduces unnecessary delay till the required number of users are found in the region. This paper proposes a two-level privacy preserving approach that effectively deals with the mentioned issues of existing techniques and provides protected services to the user with optimum level of quality of service. We suggest that before sending the exact location, perturbation is applied to the nearby cluster region at client side in order to handle the trust issues of third party. In case of single point of failure i.e. when anonymizer fails; LBS server would not be able to identify the exact location of the user due to the perturbation applied at mobile device. The proposed algorithm provides two-level location privacy protection to the user, and thus ensures smart mobility of the LBS user so that the user can freely move anywhere without any privacy breach apprehensions.

2. LITERATURE REVIEW

To handle the privacy concerns, two types of research exist in the literature. First, location cloaking based methods [3-8] and second, Private Information Retrieval (PIR) based techniques [9]. The information sent by the user (be it original or modified) should be under the control of the user who sent it [10]. Passive and Active are two different threats to the user privacy [10]. Schiller et al. [11] suggests the common architecture model of LBS with three different layers namely; positioning layer, a middleware layer and an application layer. Each layer has a dedicated responsibility in overall execution of the service.

Numerous techniques are discussed to make communication with server. One of the techniques by Dewri et al. [12] proposes a location based query in the presence of privacy supportive LBS provider. In this scheme, the user sends the query to the LBS server, even though the user uses her geographic location in a generalized way. Authors in [7] have categorized the then existing privacy preservation techniques in a hierarchical manner. Author [13] discussed the use of TTP (often called as anonymizer) as an intermediate entity that plays a key role in protecting the user's identity. Anonymizer's main aim is to hide the users' true real world identity by omitting (or modifying) the location information [14]. In the policy based scheme, the user sets the set of policies which is supposed to be followed by the service providers [13]. Due to dishonest behaviour of third party the methods that do not rely on trusted third parties are proposed [15-18]. In collaboration based method, the user do not discloses the exact location while sending a query.

The user makes changes in the location and broadcast it to her neighbours. In return, neighbours also send their modified location and centroid is calculated at user's end. This centroid value along with the user query is then sent to the LBS provider [15]. Obfuscation based methods is the process of degrading the information quality of the user's location. By using the imprecision method of obfuscation, one can easily degrade the information quality. In this method, the location space is modelled as a graph where vertices represent locations. The user sends the set of vertices instead of sending the single vertex of her own [16]. In Personal Information Retrieval (PIR) based method, the service provider cooperates with the user by following the PIR protocol, where the LBS provider answers the queries containing the location information [17]. Another relatively new approach called Privacy Enhancing Technologies (PETs) [18] restricts anonymity issues based on Trusted Computing Technologies results in a better privacy of user's personal information. Privacy Enhanced Trusted location based services (PE- TLBS) [18] focus to implement a simple protocol in which the user authenticate the server, while preserving anonymity and avoiding the possibility of their personal information leakage. The concept of dummy nodes proposes the use of dummy locations with the real location to protect the location privacy of the node [19]. The quality of requested service degrades when the number of dummy node increases.

Cloaking based approach [8][13][21] works well in protecting the user's location but vulnerable to untrusted third party i.e. Anonymizer [14], which cloaks the user's location and anonymize before sending to the service provider. Cloaking is the technique to blur the location of the node by including k-1 more nodes from the same location besides the target node [20]. In the concept of k-anonymity, locations of k users are cloaked together and all nodes in the cloak act as one of the possible sender of the query. Therefore, it becomes difficult for the third party to identify the actual user [21]. Research shows that constructing the cloak of user location does not ensure the absolute user privacy [22], however, [23] [24] propose another alternatives using TTP based and TTP free architectures.

3. PROBLEM DESCRIPTION

While cloaking the location, there is a problem of the number of users needed to form the cloak region and there may be circumstances when only single user is available for cloaking, which is surely not suitable to form the cloak. Our goal is to protect the user's private location information from the untrusted third party (or anonymizer) and at the same time mitigating the problem of minimum number of users by using the anonymizer that anonymize the location information before sending to location service provider.

3.1. Problem Formalization

We focus on the privacy of the user in a two-dimensional location based services where a trusted third party cloaks the query; Q: < location, query string >. The mechanism is protected to a great extent in the setup of dishonest third party with the problem of enough users to compute the cloak region.

4. PROPOSED ALGORITHM

4.1. Algorithm Design

In the proposed framework, cloaking region is used where the global cloaking region is split into local or sub cloak regions using clustering approach as shown in Fig 2.

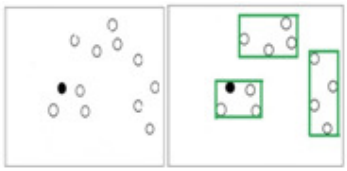


Fig 2: Cloak Region Formation: a. Global Cloak region, b. Local cloak region

The global cloak region with k users is shown in Fig 2a. The global cloak region splits into sub cloak regions as shown in Fig 2b. Dividing the global cloak region, say for k=12 users into local cloak regions for n=3; each sub local cloak region contains k'=k/n=4 users. Perturbation process is now applied to all regions before sending to the anonymizer that further anonymize the location information. Fig 3 presents the flow of proposed mechanism.

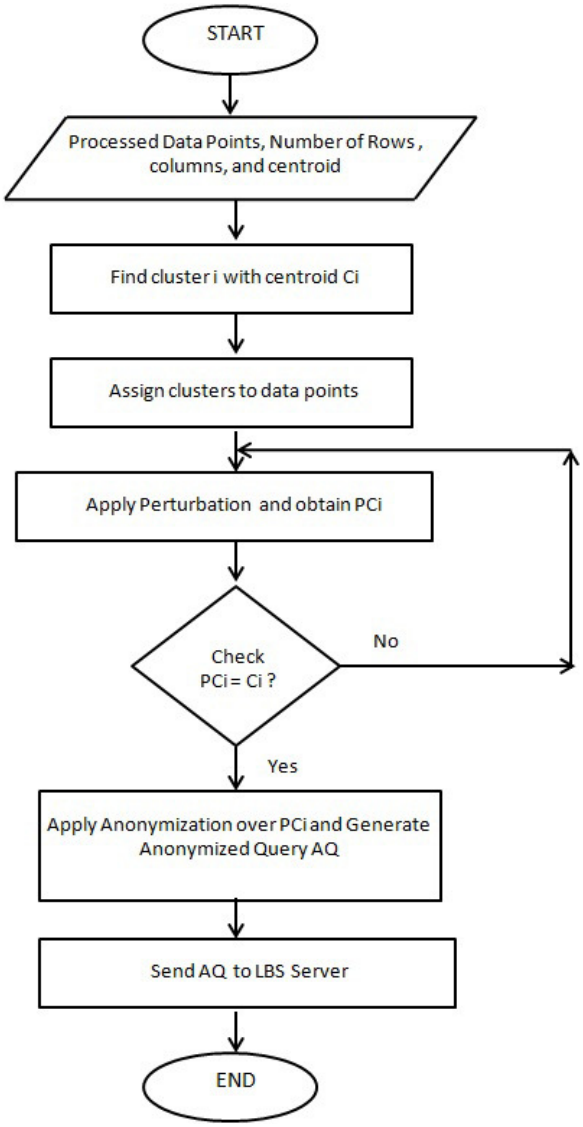


Fig 3: Flow of Proposed Mechanism

Algorithm 1:

Terminologies used:

- *NUMIT, BIGNUM* = Number of Iterations, Random Big number for minimum distance
- Mean, distort= to store coordinates sum, distorted distance
- *numRows, numCols* = Number of Rows in Input file, Number of Columns in Input file
- *numCent, newCent* = Number of Centroid in Centroid file, new centroid after Iterations
- *pCent*= Centroid after perturbation process

1. Process Input File and Centroid File
2. for(j=0 to *numRows*)
   for(k=o to *numCols*)
   mean[k]+=x[j][k] // ∑ x and ∑ y coordinates
3. for(k=0 to *numCols*)
   mean[k]/=*numRows* //calculate mean value
4. for (it=0 to *NUMIT*)
   set distort=0 ,count=0
   for(j=0 to *numRows*)
   set rmin=*BIGNUM*
   for(k=0 to *numCent*) set dist[j][k]=0
   for(e=0 to *numCols*)
   Calculate squared Euclidean distance
   END LOOP
   find Minimum Distance
   END LOOP
   distort += Minimum Distance, count++
   END LOOP
   Re-estimate new centroid points, *newCent*
   END LOOP
   Assign number of input points to respective cloak
   END LOOP
5. Call Perturbation(*newCent*);
6. Call Anonymization(*pCent*);
7. END

Algorithm 1 shows the overall functioning of the proposed algorithm. Perturbation process (step 5 of the Algorithm 1) and anonymization process (step 6 of the Algorithm 1) are described through Algorithm 2 and Algorithm 3 respectively.

Algorithm 2: Perturbation

Terminologies used:

- *numCent* = Number of centroid
- *pCent* = Perturbed centroid
- *rand( )* = Random Function

1. for(j=0 to *numCent*)
   Generate random centroid
   // *rand()%numCent*
   END LOOP
2. Assign perturbed centroid to original centroid
   pCent = *Cj;* where j Ɛ [1,j] and j is random
3. END

Algorithm 3: Anonymization

Terminologies used:

- *numRows* = Number of rows
- *pCent* = Perturbed centroid
- *temp* = holds temporary random value

1. for(j=0 to *numRows*)
   a. Generate random value and assigned to temp
   b. Anonymized point = *pCent* + temp
   END LOOP
2. Assign anonymized points to the original points
3. END

5. EXAMPLE

We have input file with coordinates {(0,2),(17,39),(10,57),(4,49),(82,31)} represented as P1, P2, P3, P4, and P5 respectively and a centroid file with coordinates {(90,10), (70,30), (50,50)} taken as C1, C2, C3 respectively. Let *numRows* be the number of rows in input file, *numCols* be the number of columns in input file and *numCent* be the number of centroid in centroid file. Here, *numRows*=5, *numCols*=2, and *numCent*=3.

Calculating mean of the given input points gives;

$\sum x = 0+17+10+4+82=113$, $\sum y = 2+39+57+49+31=178$

Therefore, mean coordinates are ($\sum x$ / numRows, $\sum y$ / numCols) i.e. (22.600000, 35.599998). Now, calculate distance matrix of each input points from each centroid will be given as,

Table 1: Distance Matrix between Centroid and Input Points

Input Points \ Centroid	C1	C2	C3
P1	8164	5684	4804
P2	6170	2890	1210
P3	8609	4329	1649
P4	8917	4717	2117
P5	505	145	1385

Dist [0][0] = $(0-90)^2 + (2-10)^2 = 8164$ {Squared Euclidean Distance}, Dist [0][1] = $(0-70)^2 + (2-30)^2 = 5684$, Dist [0][2] = $(0-50)^2 + (2-50)^2 = 4804$. Similarly, we find other distances from each input points forming the distance matrix as shown in Table 1. Now, assign input points to the nearest cluster with new centroids as; C1= (22.60, 35.60), C2= (82.00, 31.00), and C3= (7.75, 36.75). Distinctly, C2 is assigned for {(82,31)} and C3 is assigned for {(0,2), (17,39), (10,57), (4,49)}. After assigning clusters to each data point's perturbation is applied on each cluster centroid. The new perturbed centroid C2 is for input point {(0,2), (17,39), (10,57), (4,49)} and C3 is for {(82,31)}. Now, the perturbed centroid is further anonymized which resulted into the anonymized value used to contact to LBS server.

6. EMPIRICAL EVALUATION

6.1. Experimental Setup and Scenario

We implemented the algorithm on Dev-C++ version 4.9.9.2 on Intel core 2 duo 2.2 GHz machine with 4GB of RAM. We consider two entities; *McDonald's* and *Library* at five different locations. The selection of appropriate entity is based on the minimum distance between a particular Point of Interests (POIs) and entities. However, the result accuracy suffers as the number of user increases.

6.2. Results

The results of the proposed approach based on the number of iterations and number of POIs. The total computation time with respect to number of iterations and number of POIs are shown in Table 2 and Table 3 respectively.

Table 2: Time v/s Iterations, **Table 3:** Time v/s POIs

NUMIT (Number of Iterations)	Computation Time in ms
1	4.836
2	5.417
3	5.569
4	7.888
5	8.863

Number of POIs	Computation Time in ms
5	9.054
10	10.344
15	11.596
20	11.625
25	12.156
30	13.701

The above results show that as we increase the iterations, computation time of our approach increases. However, sometimes it depends on the number of processes running on the machine and may be different for different machines. It is observed that as number of POIs increases, the computation time also increases. Table 4 shows a brief comparison of our proposed approach with other existing approaches.

Table 4: Comparison with other Existing Approaches

Properties / Approaches	Accuracy	Single Point of Failure	Problem of enough users	Privacy
Mobile Device Based	✓	×	✓	✓
TTP Based	✓	✓	×	✓
Our Approach	✓	×	×	✓

Fig 4 and Fig 5 shows the computational graph of proposed mechanism with respect to Number of Iterations and Number of POIs respectively. We now present how accurate the service is provided to the requestor based on the same nearest service entity selection.

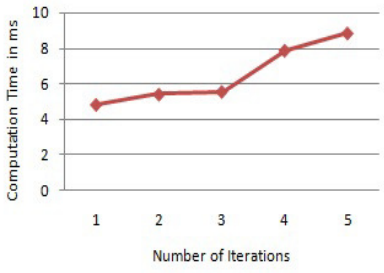


Fig 4: Time vs. Iterations

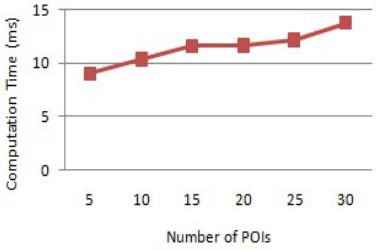


Fig 5: Time vs. POIs

Let's consider, two requested entities are McDonald's and Library with coordinates {(352.34, 534.3), (131,179.5), (192,245), (240,870), (132,564)} and {(625,387), (952,133), (287,235), (152,120), (367,755) respectively.

Now consider the input coordinates of five requesting entities i.e. P1, P2, P3, P4, and P5 according to the range in degrees as given below;

For 0-100: {(10, 12), (70, 98), (22, 87), (44, 49), (82, 31)}
For 100-200: {(110, 112), (170, 198), (122, 187), (144, 149), (182, 131)}
For 200-300: {(210, 212), (270, 298), (222, 287), (244, 249), (282, 231)}
For 300-400: {(310, 312), (370, 398), (322, 387), (344, 349), (382, 331)}
For 400-500: {(410, 412), (470, 498), (422, 487), (444, 449), (482, 431)}

Table 5: Nearest Entity Selection Table

Range	Co-ordinate Selection	
	McDonald's	Library
0-100	(131, 179.5)	(152,120)
100-200	(131, 179.5)	(152,120)
200-300	(192, 245)	(287, 235)
300-400	(352.34, 534.30)	(287, 235)
400-500	(352.34, 534.30)	(625, 387)

We also check if the above points assigned to the nearest service entity for both McDonald's and Library. Table 5 shows that which nearest service entity is selected for a particular input coordinate. For instance, P3 is assigned to both McDonald's and Library. Table 5 also shows how the same nearest service entity assigned to a particular requesting entity.

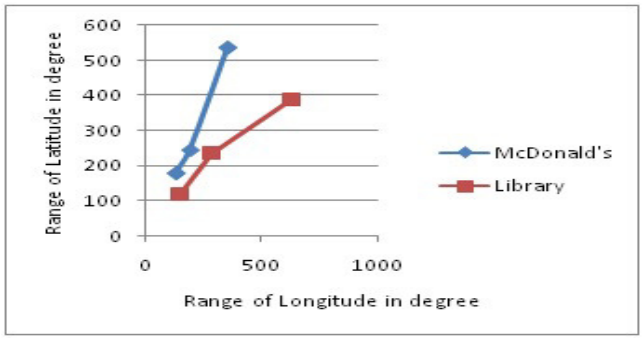


Fig 6: Nearest Entity Selection Comparison

The graph in Fig 6 shows that our approach selects same entity (McDonald's and Library) as it is selected by the original input coordinates which favours the accuracy of our approach and also provides better quality of services by selecting nearest entity in the range.

7. ANALYSIS

7.1. Privacy

Our approach is free from the issues of single point of attack and enough number of users as first problem is removed by sending the perturbed location to the trusted third party and second problem is handled by using the anonymizer which further anonymizes the location information by adding random users.

7.2. QoS

It depends on the number of users forming the cloak region. Accuracy degrades as the number of users increase i.e. k users forming the cloak achieve relatively more accurate service as compared to *k+1* users.

7.3. Communication Cost

If K is the number of clusters or cloak regions, N represents number of users in each cluster, and M is the message size then,

- Communication rounds = 4(K+1) //in general and 2(K) //if anonymizer fails
- Message size = (K+M)
- Communication Message = 4(K(M)) //in general and 2(K(M)) //if anonymizer fails

7.3. Computation Cost

In our proposed approach, time complexity can be described as *O(i*m*c*n)*. For perturbation and anonymization processes, the complexity is *O(c)* and *O(m)* respectively. Therefore, overall complexity of the algorithm is given as *O(i*m*c*n).* Where m, n represents number of rows and columns in the input file of location coordinates, i represent number of iterations, and c is the number of centroid in centroid file.

8. CONCLUSIONS

Cloaking based approaches are successful in protecting the privacy of users to some extent but there is a trade-off between the privacy and retrieved information accuracy while accessing a particular service. In order to improve the trade-off, we proposed a two-level smart privacy cloaking mechanism in which a global cloak region is divided into local regions at mobile device. After location perturbation, the location is sent to the trusted third party that brings the removal of the problem of single promising point of attack. Now, the anonymizer further anonymizes the perturbed location to handle the problem of enough users required to form the cloaking region. The proposed algorithm provides two-level location protection to the user, and thus ensures smart mobility of the LBS user so that the user can freely move anywhere without any privacy breach apprehensions. The mechanism also works well to satisfy service accuracy need of the user. Our approach is applicable for two dimensional regions and can be extended further to three dimensions; hence can provide better accuracy by covering all the scattered point of interests along with z direction.

REFERENCES

[1] B. Hofmann-Wellenhof, H. Lichtenegger, and E. Wasle, GNSS–global navigation satellite systems: GPS, GLONASS, Galileo, and more. Springer Science & Business Media, 2007.

[2] Wernke, Marius, et al. "A classification of location privacy attacks and approaches." Personal and Ubiquitous Computing 18.1 (2014): 163-175.

[3] G. Ghinita, P. Kalnis and S. Skiadopoulos, “PRIVE: Anonymous Location-Based Queries in Distributed Mobile Systems,” in Proceeding of 16th international conference on World Wide Web, pp. 371-380, 2007.

[4] G. Ghinita, P. Kalnis and S. Skiadopoulos, “MobiHide: A Mobilea Peer-to-Peer System for Anonymous Location-Based Queries,” in Proceeding of 10th International Symposium on Spatial and Temporal Databases, pp. 221-238, 2007.

[5] P. Kalnis, G. Ghinita, K. Mouratidis and D. Papadias, “Preventing Location-Based Identity Inference in Anonymous Spatial Queries,” in Proceeding of Transactions on Knowledge and Data Engineering, pp. 1719-1733, 2007.

[6] W. Ku, Y. Chen and R. Zimmermann, “Privacy Protected Spatial Query Processing for Advanced LBS,”Wireless Personal Communications 2009 Volume 51, no. 1, 2009.

[7] M. Mokbel, C. Chow and W. Aref, “The New Casper:Query Processing for Location Services without Compromising Privacy,” in Proceeding of the International Conference on Very Large Data Bases, pp. 763–774, 2006.

[8] C. Y. Chow, M. F. Mokbel, and X. Liu. A, "Peer-to-Peer Spatial Cloaking Algorithm for Anonymous Location-based Services," in Proceeding of the ACM International Symposium on Advances in Geographic Information Systems, pp. 171–178, 2006.

[9] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi and K. L. Tan, "Private Queries in Location Based Services: Anonymizers are notNecessary," in Proceeding of ACM SIGMOD international conference onManagement of data, pp. 121-132, 2008.

[10] F. Xu, J. He, M. Wright, J. Xu, "Privacy Protection in Location-Sharing Services,"in Proceeding of International Conference on Computer Application and System Modeling, ICCASM, pp. 488-491, 2010

[11] J. H. Schiller and A. Voisard, "Location-based Services: Protocol layers model," in IEEE transactions on mobile computing, pp. 29-31, 2004.

[12] R. Dewri and R. Thirumella, " Exploiting service similarity for privacy in Location based search queries," in IEEE transactions on parallel and distributed systems, vol.25, no. 2, pp. 374-383, February 2014.

[13] A. Solanas, J. Domingo-Ferrer and A. Martnez-Ballest, "Location privacy in location-Based services: Beyond TTP-based schemes," in Proceedings of the 1st International Workshop on Privacy in Location-Based Applications (PILBA), pp.12-23, 2008.

[14] L. Sweeney, "Achieving k-anonymity privacy protection using generalization and suppression, "International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, pp. 571-588, 2002.

[15] J. Domingo-Ferrer, "Microaggregation for Database and Location Privacy," in Next Generation Information Technologies and Systems, Vol. 4032, pp.106-116, 2006.

[16] M. Duckham, L. Kulik, "Location Privacy and Location-Aware Computing," in Dynamic and Mobile GIS: Investigating Changes in Space and Time, CRC Press, pp. 35–52, 2007.

[17] G. Ghinita, P. Kalnis et al., "Private queries in location based services: Anonymizers are not necessary," in Proceedingof the 2008 ACM SIGMOD international conference on Management of data, pp.121-132, 2008.

[18] S. Yan, T. F. La Porta, and P. Kermani, "A Flexible Privacy-Enhanced Location-Based Services System Framework and Practice," in Mobile Computing, IEEE Transactions on, vol. 8, no. 3, pp. 304-321, 2009.

[19] R. Kato et al., "A Dummy based anonymization method based on user trajectory with pauses," in Proceeding of 20th international conference on advances in geographic information system, pp.249-258, 2012.

[20] J M.Mano, Y. Ishikawa, "Anonymizing user location and profile information for privacy aware mobile service," in Proceeding of2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, pp. 68-75, 2010.

[21] M. Gruteser and D. Grunwald, " Anonymous usage of location based services through spatial and temporal cloaking," in Proceeding of the 1st international conference in Mobile systems, applications and services, pp. 31-42, 2003.

[22] R. Gupta and U. P. Rao, "An exploration to location based service and its privacy preserving techniques: A survey," Wireles Personal Communications, vol. 96, issue 2, pp.1973–2007, 2017.

[23] R. Gupta and U. P. Rao, "Achieving location privacy through CAST in location based services," Journal of Communications and Networks, vol. 19, no. 3, pp. 239–249, 2017.

[24] R. Gupta and U. P. Rao, "A hybrid location privacy solution for mobile LBS," Mobile Information Systems, vol. 2017, pp. 1–11, 2017.

AUTHORS

Ruchika Gupta is a Ph.D. research scholar in Computer Engineering Department, National Institute of Technology, Surat, India. Her research interests include Information Privacy, Data Security, Mobile Computing, Peer to Peer communication, and Location Privacy.

Dr. Udai Pratap Rao is currently an Assistant Professor in Computer Engineering Department at S. V. National Institute of Technology, Surat, Gujarat, INDIA. He obtained his Ph.D. degree in Computer Engineering in 2014. His research interests include Information Security and Privacy, Location Based Privacy, and Big Data Analytics.

Mr. Manish Kumar is an alumni of Computer Engineering Department, National Institute of Technology, Surat, India. His research interests include Security and Privacy in LBS.

INTENTIONAL BLANK

A PROPOSED HSV-BASED PSEUDO-COLORING SCHEME FOR ENHANCING MEDICAL IMAGES

Noura A. Semary

Faculty of Computers and Information, Menoufia University, Egypt

ABSTRACT

Medical imaging is one of the most attractive topics of image processing and understanding research fields due to the similarity between the captured body organs colors. Most medical images come in grayscale with low contrast gray values; which makes it a challenge to discriminate between the region of interest (ROI) and the background (BG) parts. Pseudo-coloring is one of the solutions to enhance the visual appeal of medical images, most literature works suggest RGB-base color palettes. In this paper, pseudo-coloring methods of different medical imaging works are investigated and a highly discriminative colorization method is proposed. The proposed colorization method employs HSV/HSI color models to generate the desired color scale. Experiments have been performed on different medical images and different assessment methods have been utilized. The results show that the proposed methodology could clearly discriminate between near grayscale organs especially in case of tumor existence. Comparisons with other literary works were performed and the results are promising.

KEYWORDS

Medical imaging, pseudo-coloring, colorization, HSV

1. INTRODUCTION

Medical images not only represent structural appearance information, they are also capable of examining complex and sophisticated internal biological processes. Medical imaging contributes to many disease diagnoses and also plays an important role in understanding the human anatomy which guides surgical assistance during the procedure.

Medical Imaging has various modalities and applications. X-ray, CT, MRI, Ultrasound, Mammogram, Nuclear Medicine, PET, Ultrasound, and Thermal imaging are some of the famous imaging technologies [1, 2]. Each of which has its suitable applications and features and exports gray shades images which usually come in low contrast intensities.

Digital image can be represented in different formats; 1) Grayscale (8-bits/pixel), 2) True Color (24 bits/pixel) and 3) Indexed (8-bits/pixel index image + Color Map) [3]. Medical images usually come in grayscale which has only 256 gray shades variations. While most image understanding researchers prefer to deal with color images instead of grayscale, as the color variations exceed 16 million degrees of colors.

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 81–92, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80407

For some imaging technologies, a pseudo-coloring system may be embedded in the imaging device. Pseudo-Coloring gives non-real colors to the grayscale image by converting it to an indexed image with a fixed color map. Generating the color map is the main contribution in this field of research.

There are two approaches for medical images colorization in the literature; semi-automatic (interactive) pseudo-coloring [1] and automatic (non-interactive) pseudo-coloring [4-10].

In this paper, we are concerned with automatic medical images pseudo-coloring. The rest of this paper is organized as follows, Section 2 presents automatic medical colorization literature review. The proposed colorization method is presented in Section 3 in details. Section 4 presents the experimental results and a comparative study of the proposed method and other different methods. Finally, Section 5 concludes this paper.

2. LITERATURE REVIEW

L. H. Juang and M. N. Wub [4] used color-converted segmentation with K-means clustering technique for Brain MRI tumor objects colorization and tracking. Starting with a gray image, the original image was segmented by k-means and mapped colors to the segments by using color-convert which starts by R, G, B then mapped to a single index value. Compare between Otsu segmentation results and claim 96% accuracy and 10 minutes processing time!

M. d. C. V. Hernández et. al. [5] used image fusion between T2W and FLAIR images for differentiating normal and abnormal brain tissue, including white matter lesions (WMLs). They modulated two 1.5T MR sequences in the red/green color space and calculated the tissue volumes using minimum variance quantization. M. Attique et. al. [6] also used image fusion for brain MRI images. They utilized Single slice of T2-weighted (T2) brain MR images using two methods; (i) A novel colorization method to underscore the variability in brain MR images, indicative of the underlying physical density of bio-tissue, (ii) A segmentation method to characterize gray brain MR images.

M. Martinez et. al. [7] proposed their color map for CT Liver images, the ROI is selected manually. Each grayscale pixel was assigned a color value (R, G, B) based on a generated color map. A color scheme was developed where the lowest tissue density value was colored red, blending towards green as the tissue density value increases and continued to blend from green to blue for the next range of increasing tissue densities. An associated segmentation process is then tailored to utilize this color data. It is shown that colorization significantly decreases segmentation time. M. E. Tavakol et. al [8] applied their proposed colorization system on Thermal Infrared Breast Images. Lab color model is considered. Two color segmentation techniques, K-means and fuzzy c-means for color segmentation of infrared (IR) breast images are modelled and compared.

Z. Zahedi et. al [9] proposed their colorization system for breast thermal images as a nonlinear function transforms for pseudo-coloring of infrared breast images based on physiological properties of the human eye. N. S. Aghdam et. al [10] proposed four pseudo-coloring algorithms for breast thermal images. The first two algorithms are in HSI color space and the other two are in CIE L*a*b*.

3. PROPOSED COLORIZATION SCHEME

The proposed colorization system is based on HSV/HSI color models where any RGB color triple value can be expressed in terms of Hue (H) Saturation (S) and Intensity (I or value V). Since the grayscale image has only intensity component and has no hue or saturation values [3], our system is based on generating suitable hue and saturation values for each intensity level. Figure 1 presents the main block diagram of the proposed system.

3.1. Intensity Component Generation

In order to generate a carefully designed pseudo-color coding which could preserve all the information of grayscale images and does not generate any distortion in the image, the grayscale image is loaded and saved as the intensity component for the output image.

$$I = G \tag{1}$$

, where L is the Intensity component in HSV/HIS image and G is the gray image.

3.2. Hue Component Generation

Since the gray image has different gray shades which reflect different medical meanings, it's suggested to use the same varieties for the domain color of each region. Human vision can discriminate between the main Rainbow colors; red, orange, yellow, green, cyan, blue and magenta, which can be generated by Hue component. Hue component reflects the dominant color of the pixel. Here, we studied 3 strategies of generation the Hue component.

A. Equal to Intensity (EI)

In this strategy, Hue is set to the same value of Intensity (2). That makes the shades vary from red to magenta for the gray shades from black to white respectively. For some medical images, the ROI gray shades are in light areas with very close values. By using this strategy the colors of ROI may come in red and magenta which may not be discriminative enough.

$$H_{EI} = G \tag{2}$$

B. Complement of Intensity (CI)

In this strategy, Hue is set to the Intensity complement value (3). That makes the shades vary from red to magenta for the gray shades from white to black respectively

$$H_{CI} = 1 - G \tag{3}$$

C. Stretched Inverted Intensity (SI)

Since the gray levels of medical images are very close, a stretched grayscale image has been generated by contrast stretching function (4). The aim of stretching function is to generate wider color space than the limited gray levels in the source gray image

$$H_{SI} = 1 - \frac{(b-a)}{(d-c)}(G-c) + a \quad (4)$$

, where G is the input gray image. The parameters a and b are the minimum gray level and the maximum gray level in the input image, while c and d are the minimum gray level and the maximum gray level in the desired image respectively.

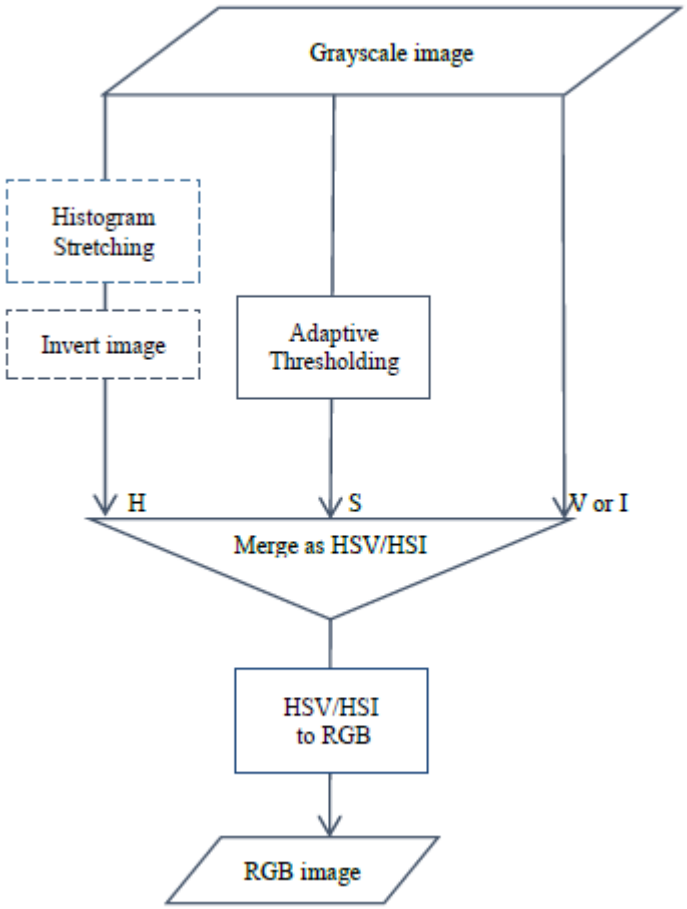


Figure 1. Proposed colorization system

3.3. Saturation Component Generation

For the Saturation component, the original gray image is used again to generate the suitable saturation. Since it is important to see the organs in full clear color, while neglecting the background, an adaptive thresholding [3] has been used to generate a binary image of black background pixels and white foreground pixels (5).

$$S = \begin{cases} 1 & I \geq th \\ 0 & I < th \end{cases} \quad (5)$$

The complete process can be illustrated mathematically by (6).

$$HSL = \begin{cases} H_{SI} = 1 - \frac{(b-a)}{(d-c)}(G-c) + a & \\ S = 1 & I \geq th \\ S = 0 & I < th \\ L = G & \end{cases} \quad (6)$$

There is a little difference between HSV and HSI color models; where the pure colors of full saturation are at the top of HSV while they are in the middle at the HSI. In our proposed system, both models are used considering the same transformations (6). Finally, the final RGB color image is generated after the well-known HSV/HSL to RGB conversion [3].

4. EXPERIMENTAL RESULTS

The proposed system has been implemented in MatlabR2013 on 8G-RAM 64bit-OS Windows 8 machine and different medical imaging modalities have been used for testing. A comparison between the proposed coloring system and other systems has been made with regard to different assessment methods.

4.1. Subjective Assessment

Subjective assessment methods are utilized to perform the task of assessing visual quality to the human subjects. This section presents the visual results of the proposed system as well as different literature methods [7-10]. Figure 2.a presents a grayscale image that presents the 256 shades of gray and the color palettes for the methods of [7-10] compared to the proposed method with 6 strategies; (3 for HSV and 3 for HSL). In this example, the saturation threshold has been set to 0.1 for illustration. Figure 3 presents more visual results for different types of medical imaging; Brain MRI, Breast Thermal, Liver CT, and Bone MRI. Generally, it seems that our proposed system could successfully discriminate between the background and the foreground, since it gave colors only to the desired imaged organs. It appears that it gives more details to the images by giving a lot of color verities from hot (red) to cold (blue).

Mean Opinion Score (MOS) [11] is a subjective measure that can be presented in numbers. It measures the users' satisfaction with the visual results. It is calculated by averaging the results of a set of subjective test that gives a score to the results image from 1 (bad) to 5 (Excellent). The mean rate of a group of observers who join the evaluation is usually computed by the following equation:

$$MOS = \sum_{g=1}^{5} gp(g) \quad (7)$$

, where g is grade and p(g) is grade probability. We have performed the MOS test using an online test. 60 volunteers have rated the different 12 methods we presented in this paper by giving them rates from 1 to 5. The obtained MOS is presented in Figure 4.

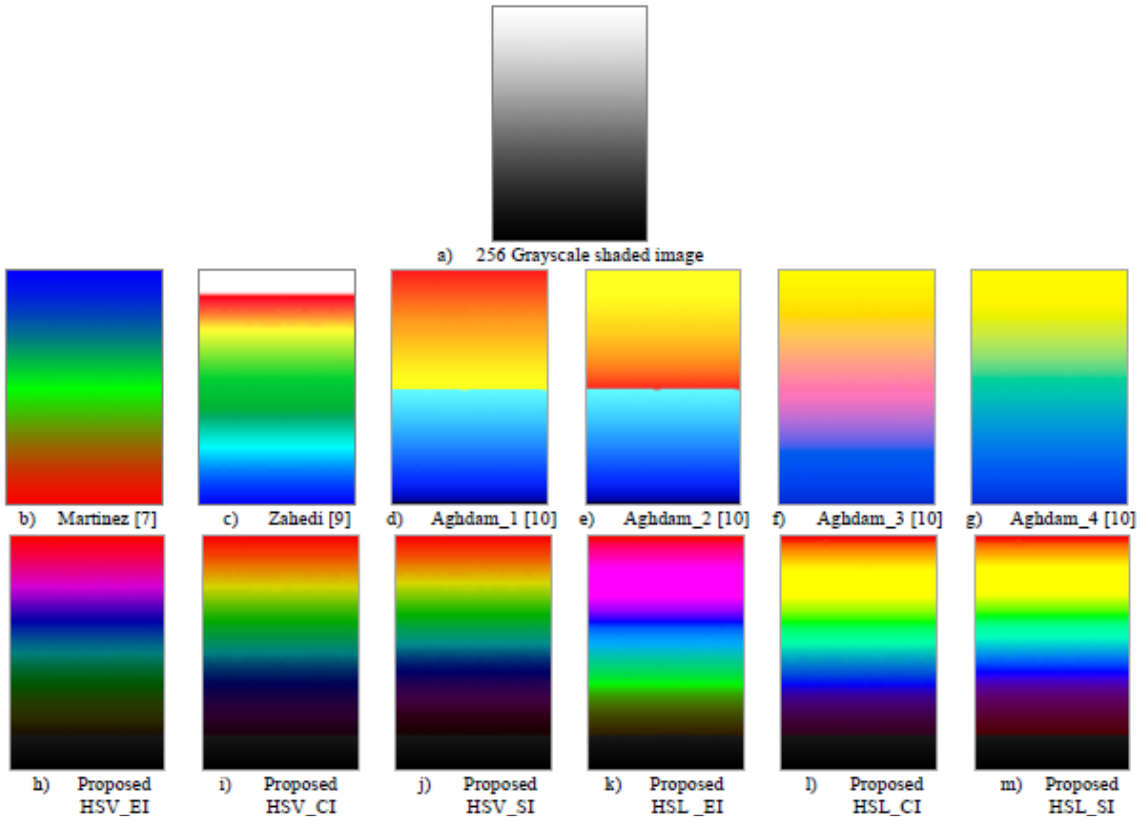
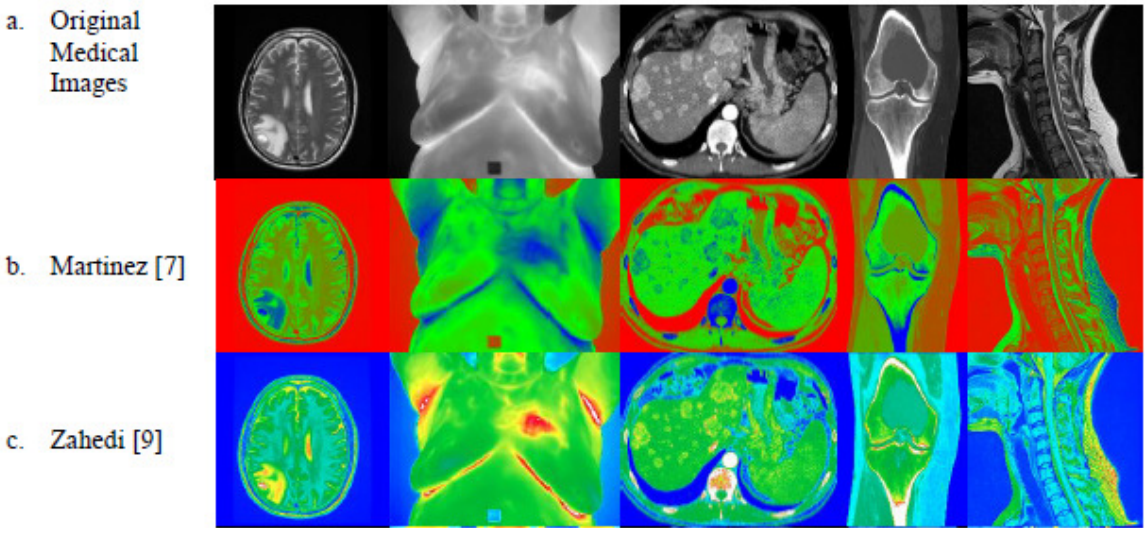
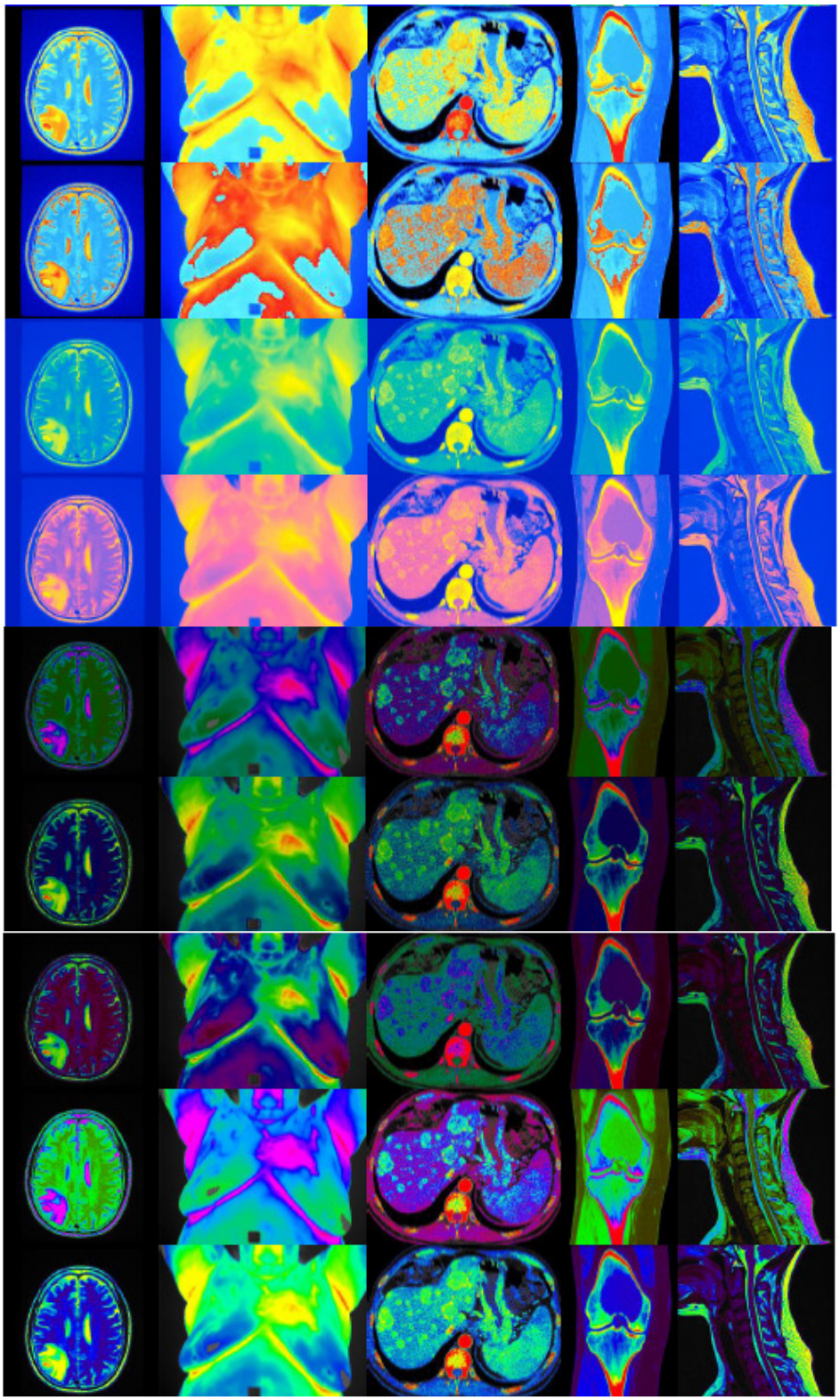


Figure 2. (a) 256 shaded gray palette (b-g) Color Palettes of [7-10] methods and (h-m) the proposed color palette with 6 strategies.





d. Aghdam HSI [10]

e. Aghdam Lab [10]

f. Proposed HSV-EI

g. Proposed HSV-CI

h. Proposed HSV-SI

i. Proposed HSL-EI

j. Proposed HSL-CI

k. Proposed HSL-SI

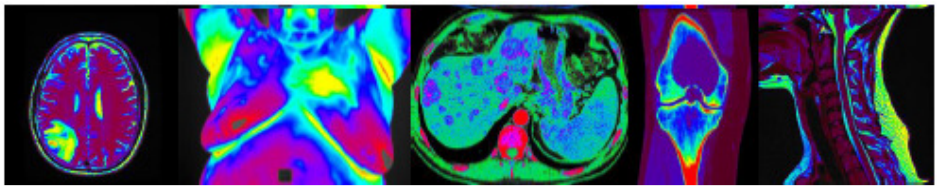


Figure 3. (a) Original medical gray image (b-e) Visual comparison between literature methods [7-10] and (f-k) the proposed colorization method with 6 strategies

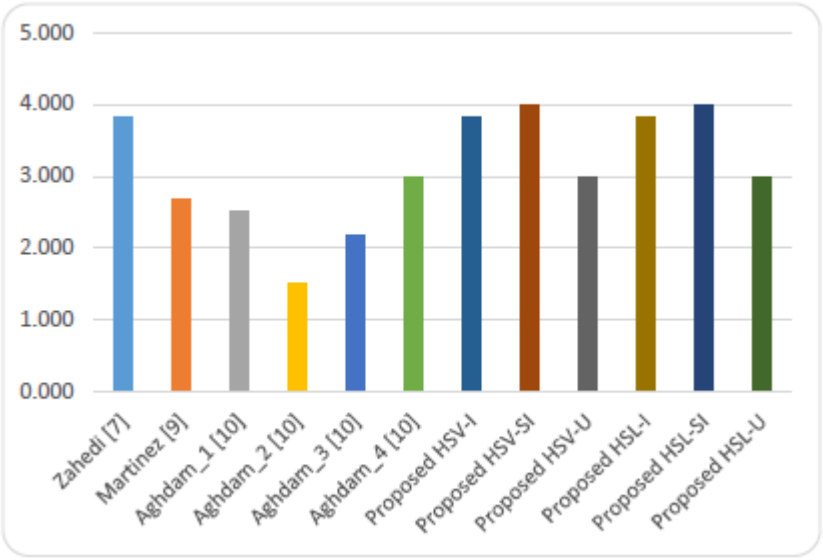


Figure 4. MOS for methods in [7-10] and the proposed color palettes

4.2. Objective Assessment

It's difficult to find the best objective metric to evaluate gray to color images conversion since no reference color image is provided. MSE (8) and PSNR (9) are widely known objective quality assessment methods for image enhancement. Since they measure the difference between two images in the same color space, it makes no meaning to get the MSE between a gray image and a colored one. Some works consider the MSE as an objective assessment method to their colorization methods by either getting the MSE between the original color image and the recolored one or to convert the gray image into RGB image with equal R, G and B values and calculate the difference between the gray RGB image and the colored one. In the latter case, it seems the higher the MSE, the higher the distance between color and gray values, the best the colorization result is.

$$MSE = \frac{1}{3MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[G(i,j) - Cl(i,j)]^2 \quad (8)$$

$$PSNR = 10\,log\left(\frac{255^2}{MSE}\right) \quad (9)$$

, where G is the gray image with 3 channels; R, G and B, and Cl is the colored image in RGB space.

Normalized Color Difference (NCD) [3] [12] is another objective measure was used by some researchers. NCD represents the distances between colors in a given color space (Usually Lab color space). The larger the NCD, the worse the image quality is. The NCD indicator is calculated using the following formula:

$$NCD = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{\sum_{q\in[L,a,b]}\left[G_q(i,j)-Cl_q(i,j)\right]^2}}{\sum_{i=1}^{M}\sum_{j=1}^{N}\sqrt{\sum_{q\in[L,a,b]}Cl_q(i,j)^2}} \quad (10)$$

, were G_q and Cl_q are the gray and the colored images with q channel (in Lab space).

Structure Similarity Metric (SSIM) [13, 14] is another objective measure that was proposed based on the human visual system. SSIM reflects how the colorization process affects the structure of the image. Popowicz [15] improved the SSIM by adding a color comparison to the criteria of the grayscale SSIM. Since the SSIM is defined only for grayscale images, it can be adapted for color image colorization by calculating the SSIM for every single color channel independently, then calculating the mean. When taking into consideration decorrelated color spaces, where the channels are orthogonal, we may assess the quality with the root of the sum of squared SSIM results obtained for each channel. The detailed Mean SSIM (MSSIM) could be found in [15].

In this work, MSE, PSNR, NCD, and SSIM measures have been calculated for works in [7-10] as well as the proposed method. Table 1 presents the objectives measures calculated for the proposed palette as well as the methods in [7-10]. As discussed before, MSE and PSNR give different meaning when it comes to colorization. As it is extremely sensitive to the offsets in color or luminance channels. It is possible that, although the visual impression of the image does not change, the PSNR value may be much lower, indicating the poor quality. From the table, it seems the proposed palettes have high MSE and low PSNR. Since NCD is the most suitable measure for colorization procedures assessment [15] it is considered for assessing the 12 methods. The lower the NCD, the better the image quality. Since our proposed system keeps the luminance channel untouched, it's expected to have high MSSIM index. As the MSSIM could be presented as a grayscale image, Figure 5 presents the MSSIM for the Brain MRI image from our test set (Figure 3).

5. CONCLUSION

In this paper, we propose HSV/HSL pseudo-coloring schema for medical images. We have proposed 6 strategies based on the basic proposed structure. The proposed strategies vary in colors range (from red to blue and magenta or vice versa). Using HSV/HSL models enable giving rainbow colors to the gray shades either in the same order or with an inverse order. We recommend using the same gray image as Intensity for keeping the same structure of the image as well as the original intensity information. The recommended binary Saturation channel gives clear pure colors to the captured organs while neglecting the black background here what enhances the visual appeal of the images. Also, it increases the color difference between the original grayscale image (where saturation is set to 0) and the colored image (where the saturation is set to 1). For Hue component generation, three strategies have been suggested; Equal to Intensity (EI), Complement of Intensity (CI) and Stretched Inverted Intensity (SI). Hue contrast stretching has been recommended for giving more colors to a smaller range of gray shades. Subjective and objective testing has been performed showing that the proposed system has high

MOS and SSIM, and low NCD. We can conclude that using HSV/HSB color models in medical images is a good choice subjectively and objectively.

Table. 1. Objective measures for literature palettes [7-10] and the proposed palettes

	Color Space	MSE	PSNR	NCD	SSIM
Martinez [7]	RGB	1.19E+04	16.9628	0.5497	0.7589
Zahedi [9]	RGB	5.09E+03	25.4766	0.4488	0.7874
Aghdam [10]	HSV_1	6.90E+03	22.4306	0.4803	0.8195
	HSV_2	5.30E+03	25.0627	0.4303	0.8502
Aghdam [10]	Lab_1	4.62E+03	26.4341	0.3713	0.8562
	Lab_2	5.99E+03	23.8435	0.3854	0.9834
Proposed	HSV-U	1.06E+04	18.178	0.2933	0.9721
	HSV-SI	1.05E+04	18.2176	0.297	0.9454
	HSV-I	1.06E+04	18.178	0.3366	0.9721
	HSL-U	8.88E+03	19.9045	0.3302	0.9186
	HSL-SI	8.86E+03	19.9339	0.3247	0.8883
	HSL-I	8.88E+03	19.9045	0.3907	0.9186

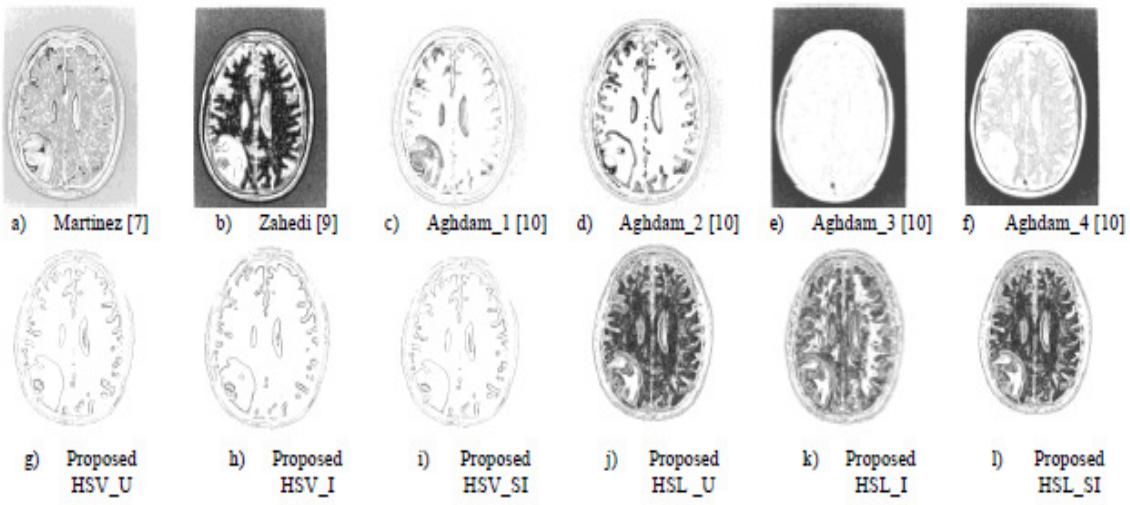


Figure 5. (a-f) MSSIM for Color Palettes of [7-10] methods and (g-l) MSSIM for the proposed color palettes.

REFERENCES

[1] Lagodzinski, P., & Smolka, B. (2009, October). Colorization of medical images. In Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference (pp. 769-772). Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee.

[2] Khan, M. U. G., Gotoh, Y., & Nida, N. (2017, July). Medical Image Colorization for Better Visualization and Segmentation. In Annual Conference on Medical Image Understanding and Analysis (pp. 571-580). Springer, Cham.

[3] Plataniotis, K. N., & Venetsanopoulos, A. N. (2013). Color image processing and applications. Springer Science & Business Media.

[4] Juang, L. H., & Wu, M. N. (2010). MRI brain lesion image detection based on color-converted K-means clustering segmentation. Measurement, 43(7), 941-949.

[5] Hernández, M. D. C. V., Ferguson, K. J., Chappell, F. M., & Wardlaw, J. M. (2010). New multispectral MRI data fusion technique for white matter lesion segmentation: method and comparison with thresholding in FLAIR images. European radiology, 20(7), 1684-1691.

[6] Attique, M., Gilanie, G., Mehmood, M. S., Naweed, M. S., Ikram, M., Kamran, J. A., & Vitkin, A. (2012). Colorization and automated segmentation of human T2 MR brain images for characterization of soft tissues. PloS one, 7(3), e33616.

[7] Martinez-Escobar, M., Foo, J. L., & Winer, E. (2012). Colorization of CT images to improve tissue contrast for tumor segmentation. Computers in Biology and Medicine, 42(12), 1170-1178.

[8] Tavakol, E. M., Sadri, S., & Ng, E. Y. K. (2010). Application of K-and fuzzy c-means for color segmentation of thermal infrared breast images. Journal of medical systems, 34(1), 35-42.

[9] Zahedi, Z., Sadri, S., & Moosavi, A. (2012, December). Breast thermography and pseudo-coloring presentation for improving gray infrared images. In Photonics Global Conference (PGC), 2012 (pp. 1-5). IEEE.

[10] Aghdam, N. S., Amin, M. M., Tavakol, M. E., & Ng, E. Y. K. (2013). Designing and comparing different color map algorithms for pseudo-coloring breast thermograms. Journal of Medical Imaging and Health Informatics, 3(4), 487-493.

[11] George, A. G., & Prabavathy, K. (2014). A survey on different approaches used in image quality assessment. International Journal of Computer Science and Network Security (IJCSNS), 14(2), 78-84.

[12] Popowicz, A., & Smolka, B. (2017). Fast image colourisation using the isolines concept. Multimedia Tools and Applications, 76(14), 15987-16009.

[13] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

[14] Wang, Z., Bovik, A. C., & Sheikh, H. R. (2005). Structural similarity-based image quality assessment. Digital Video image quality and perceptual coding, Series in Signal Processing and Communications, H. R.Wu and K. R. Rao, Eds. CRC, 225-241.

[15] Popowicz, A., & Smolka, B. (2015). Overview of Grayscale Image Colorization Techniques. In Color Image and Video Enhancement (pp. 345-370). Springer International Publishing.

AUTHORS

Noura A. Semary works as an Assistant Professor in Faculty of Computers and Information, Menoufia University, Egypt. She has B.Sc. in 2001 from Cairo University, Faculty of Computers and Information. She worked as a Staff Member in the Faculty of Computers and Information, Menoufia University, Egypt in 2003. In 2007 and 2011 she has obtained her Master and Ph.D. degrees respectively in Information Technology from Computers and Information Faculty at Menoufia University. Her main research interests are in Image Processing, Computer Vision, Data Compression, Data Hiding, Virtual Reality and Assistive Technology fields. In 2009 she won the first rank in “Made in Egypt” and “Made in Arab World” competitions for her project “Black and White Movies Colorizer”. She has about 40 publications in international conferences and journals.

AN ONTOLOGY-BASED DATA WAREHOUSE FOR THE GRAIN TRADE DOMAIN

Mhamed Itmi[1] and Boulares Ouchenne[2]

LITIS Laboratory, INSA Rouen Normandy,
University of Rouen Normandy. France

ABSTRACT

Data warehouse systems provide a great way to centralize and converge all data of an organization in order to facilitating access to the huge amounts of information, analysing and decision making. Actually, the conceptual data-models of data warehouses does not take into account the semantic dimension of information. However, the semantic of data models constitute an important indicator to help users to finds its way in any applications that use the data warehouse. In this study, we will tackle this problem trough using ontologies and semantic web techniques to integrate and model information. The contributions of this paper are an ontology for the field of grain trade and a semantic data warehouse which uses the ontology as a conceptual data-model.

KEYWORDS

Ontology Engineering, Data Warehouse, Sparql Endpoint.

1. INTRODUCTION

HAROPA Port of Rouen has a leading position in France with around 50% market share of wheat exported by sea and 44% on barley. According to the highly dependent campaigns of the production and the meteorological conditions, the grain traffic represents in tonnage between 25% and 30% of the traffics of Rouen.

HAROPA port of Rouen benefits from the proximity of a market of 22 million consumers in a radius of 200kms. The expertise in transport, handling and logistics of its operators, combined with its geographical location, explain its strategic interest for all types of goods. It is mainly known for the export of cereals.

With the aim of consolidating HAROPA's leadership as a major European port for the export of cereals and to strengthen its competitiveness and to capture new markets for the actors of the cereals sector of the Seine axis, we have participated in a study which led us to work on an ontology dealing with the grain trade.

Nowadays, organization's data sources are scattered across multiple systems, not necessarily compatible. These data sources are designed to be effective for the functions on which they are specialized. They are of- ten unstructured for analysis and designed with the primary objective of preserving information. As critical business information has to be served with a fast response time and well structured for decision making. The data warehouse aims to aggregate and enhance data from different sources to allow the user to get access easily, quickly and ergonomically to the

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 93–101, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80408

information. The process of implementing a data warehouse is a very complex task that pushes designers to acquire wide knowledge of the domain, thus requiring a high level of expertise. The design of the conceptual model is the key step of the process of designing data warehouses. This model is the basis for the implementation of the data warehouse. The conceptual model of a data warehouses is generally [1,2,3] represented by a standard 3NF data model, star models, snowflake model, or constellation model. These models prescribes the information to be represented in a database stored on a physical medium. These data models are only powerful at the structural level and lack the ability to specify the semantic relations contained in complex data for modelling and analysis. To tackle this problem, we will use ontologies to facilitate the integration of heterogeneous data sources by resolving semantic heterogeneity between them [4,5,6]. The main advantages of using ontologies is to define the semantic vocabulary of data and to obtain implicit knowledge thanks to performing reasoning on it.

The remaining part of the paper is structured as follows: section 2 introduces the industrial case study including and describes the engineering process performed to develop our ontology. Section 3 presents some of the content of our ontology. Section 4 describes the architecture of our proposed model for the data warehouse, and summarizes implementation we have conducted. Section 5 summarizes the work and draws conclusions.

2. THE CONTEXT OF THE STUDY AND THE GENERAL APPROACH

The case study presented in this paper is based on an industrial research and development project. The goals of this study is to design a reference ontology for representing information on the grain trade activity at the port of Rouen. This ontology will serve as a data model for a data warehouse containing information collected as part of an R&D project. There are many reasons for us to adopt the choice of implementing a data warehouse:

1. Centralize and converge all types of data collected from several for- mats (reports, databases, Excel files, etc.) to semantic data interlinked in order is facilitate access to information, analysis and decision making.

2. Allows a balanced perspective of the organizations. Indeed, insofar as each relevant indicator (for example juridical) is directly or indirectly correlated to another (for example economic), it affects directly or indirectly its objectives.

3. Saves time and money. In fact, users can quickly access data from a huge number of sources (all in one place). So, they can quickly make the best decisions.

4. Improves the quality and consistency of data. As its implementation includes the conversion of data from many sources in a common format. So we can have more confidence in the accuracy of such data.

5. Provides historical information. Indeed, it can store large amounts of historical data so we can analyse the different periods and temporal trends in order to make predictions.

2.1. The General Approach

Figure 1 shows an overview of our approach. The first step in the design of a data model of the warehouse. Once the model is developed and validated, the second phase begins. It consists on the creation of the data warehouse as a dataset (RDF [7] database). The third and final step in the population of the data warehouse. It takes place in several steps and constitute the data migration

phase after they have undergone selection and reformatting operations in order to be homogenized.

This phase is an important step insofar as it is estimated at about 60 per cent of the implementation time of the warehouse. We have identified several types of data sources to populate the warehouse. For each type of data we have implemented integrators and, for the interaction between users and the data warehouse, we developed web interfaces for navigation and update.

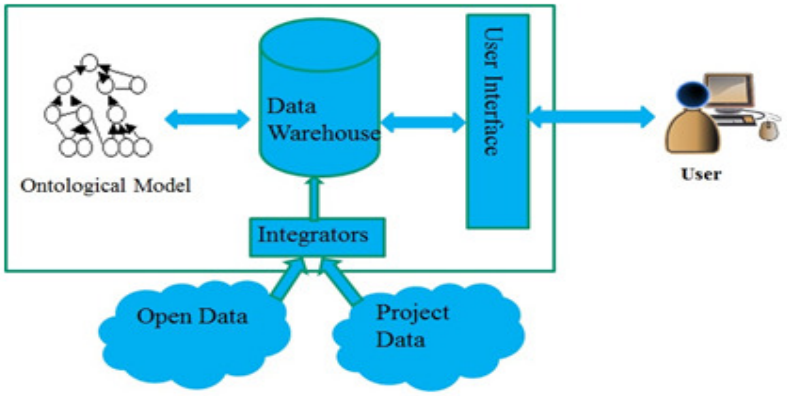


Figure 1. Overview of the approach

3. OVERVIEW OF THE ONTOLOGICAL MODEL

This starting point of the process of the ontology development is to define its domain and scope. In our case, the domain and scope of the ontology is the grain trade activity at the port of Rouen. The process for developing our ontology follows an iterative and incremental approach [8,9]. We have followed the following steps:

Step 1: Creating a conceptual data model that identifies and structure the basic concepts implemented by our ontology. The task of Conceptual Modelling plays a crucial role in the process of our ontology development. Conceptual models translate and specify the main data requirements in an abstract representation about our domain.

Step 2: Identifying pre-existing RDF or OWL schemas which propose classes and properties equivalent to those identified in the previous step to expand and/or refine them. Reusing existing ontologies, may even be a requirement if our data warehouse needs to interact with other applications already use specific ontologies or controlled vocabularies.

Step 3: Enriching the ontology by adding features to certain relation-ships, such as the fact that a property is transitive, reflexive, symmetric, etc.

During the design of this ontology, we have strived apply the commonly recommended techniques in the Community [10]. For instance, the definition for each object property a reverse property thus facilitating the manipulations and aligning our ontology with other reference ontologies (such as the FOAF Ontology [11])). Also, we opted for reusing only terms we need from external ontologies, without importing them explicitly. Finally, our ontology has been described in OWL2, we have taken advantage of possibilities of this language in terms of expressiveness [12]. In particular, we defined chains of properties to infer new relationships without recourse to a language dedicated to the expression rules.

3.1. Overview of the Ontological Model

The model we have designed includes all information relating to the activity of grain trading at the port of Rouen. It represents graphically the entities, the various actors and contractual relations between them (Sale, Rent, Fobbing, etc.). In the remainder of this article, each part of our ontology is presented as an UML class diagram where:

- UML classes are OWL classes.
- UML class attributes represent OWL data properties (data types have been added to not to complicate the presentation).
- Association relationships between OWL classes represent OWL object properties.
- Each identifier is prefixed by a domain name. Only those prefixed by (realgrain) are actually introduced in our ontology and he others are reused from other ontologies.

To be readable, our ontological model is divided into three parts: the actors, the different types of contracts and the various kinds of sale contracts.

3.1.1. Representation of Actors

This section describes the different information describing people, organizations and infrastructure involved in the export of grain field, the nature of that involvement and the semantic relationships between them. Figure 2 shows the class diagram of this first part of the ontology. For instance, considering the example of a Silo which is a structure used to store the grain before its shipments to final clients. The Silo class defined in our ontology is a subclass of

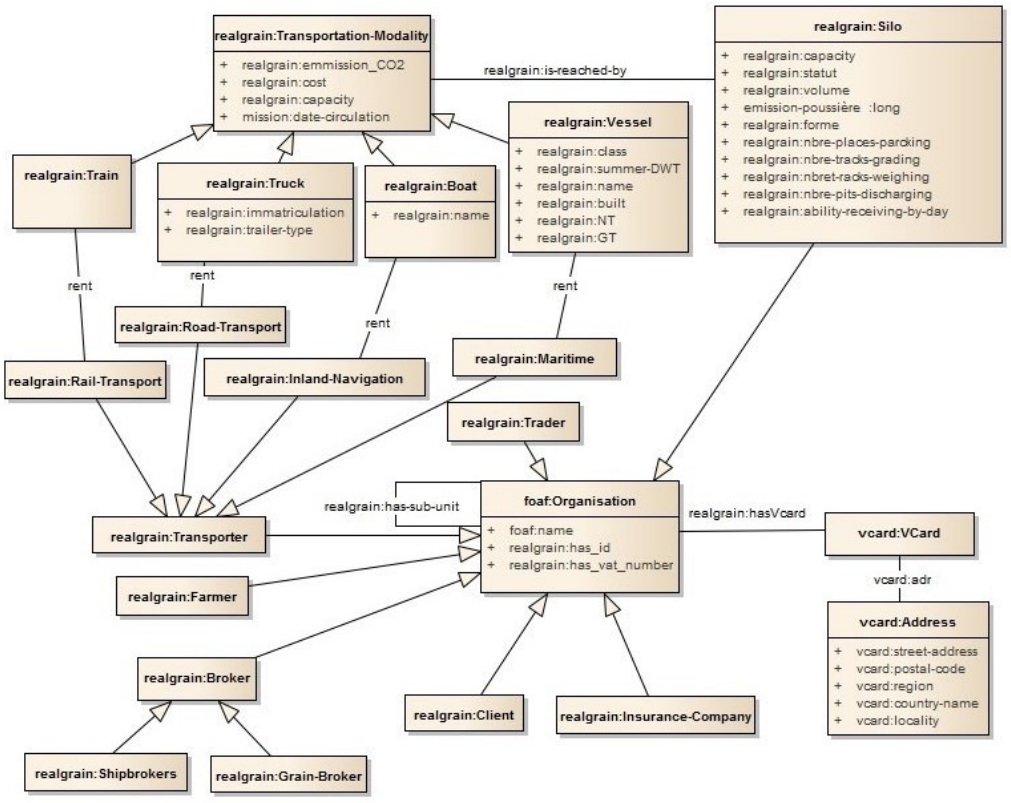


Figure 2. Actors

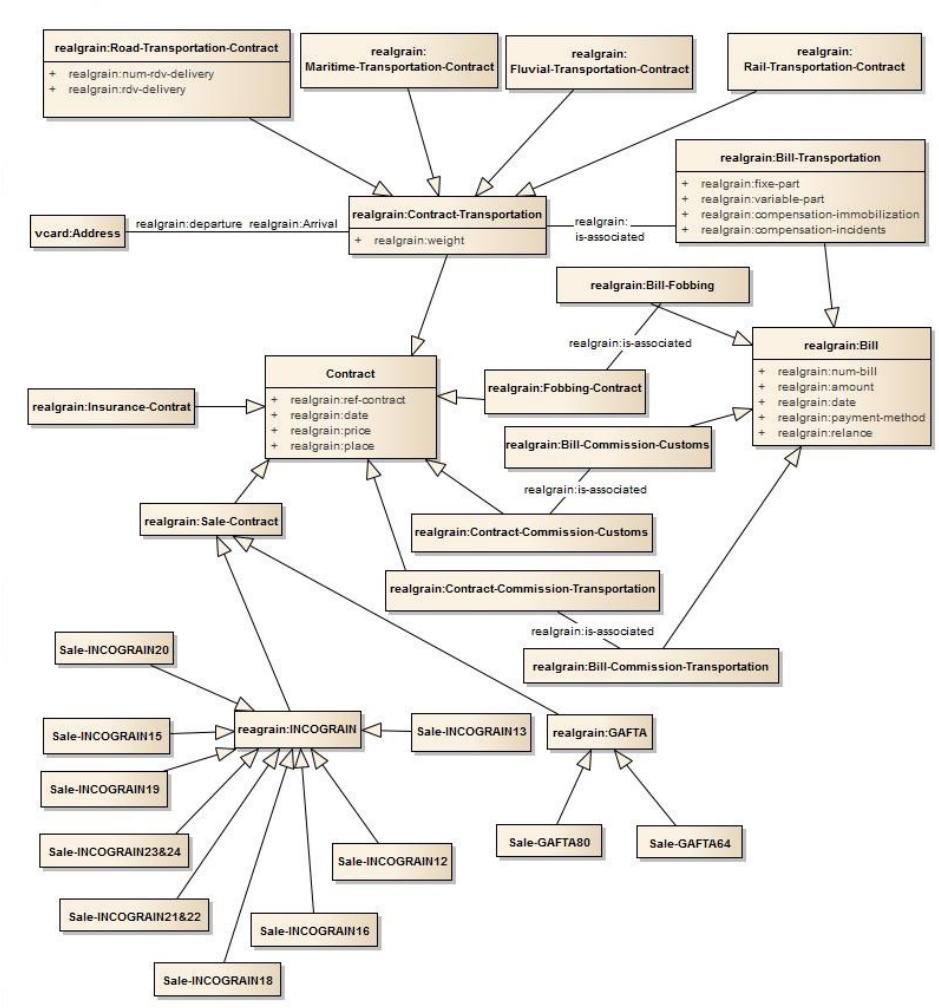


Figure 3. Contracts

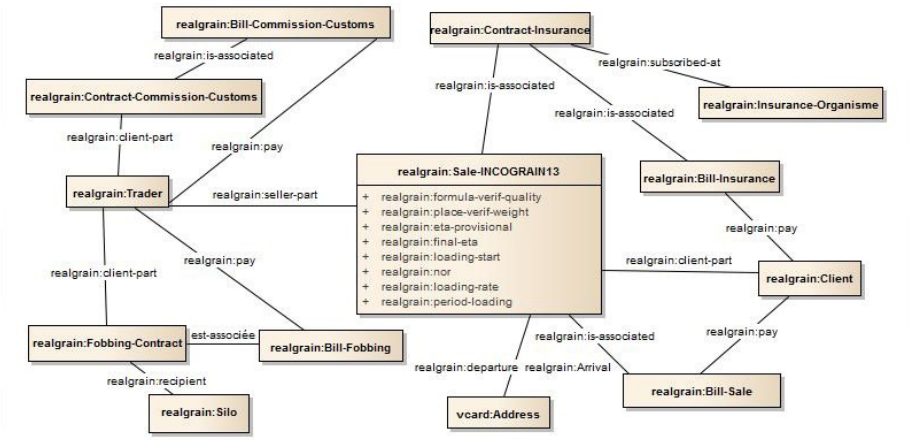


Figure 4. Example of an INCOGRAIN Contract

the foaf:Organization class defined in the FOAF ontology [11]. We have introduced and reused some data properties describe relationships between instances (individuals) and data values. We have also introduced and reused some Object properties to describe relationships between two instances (individuals). They link individuals from a domain to individuals a range. For example:

- foaf:name: is a data property reused from the FOAF ontology to represent the name of the organization.
- realgrain:capacity: is data property introduced to represent storage capacity of the Silo.
- vcard:adr: is an object property reused from the Vcard ontology to link the postal address to an information contact.
- realgrain:is-reached-by is an object property introduced to represent the different kinds of transportation modalities to reach the Silo.

3.1.2. Contracts

This section describes the different contracts, the information which contains and semantic relationships between them. Figure 3 shows the class diagram of the second part of ontology, and the hierarchical relationships between the different classes. Considering the example of sale contracts. For export sales of French cereals, buyers and sellers use standard contracts of purchase and sale. There are mainly two kinds of contracts GAFTA and INCOGRAIN which facilitate transactions and reduce sources of disputes between buyers and sellers. The choice of one of them determines the place of arbitration (Paris and London), the language of the proceedings (French or English) and applicable law (French or Anglo-Saxon). For The INCOGRAIN, there are twelve contract type available in several languages. Each contract is identified by a number which determine the client, the seller, the mode of transportation used for bringing the grain, etc.

3.1.3. Sales Contracts

For international sales of cereals from the port of Rouen, there are two agreement models of contracts CAF and FOB. The main difference between an FOB and a CAF agreement is the point at which responsibility and liability transfer from seller to buyer. With a FOB shipment, this occurs when the shipment reaches the port or other facility designated as the point of origin. With a CAF agreement, the seller pays costs and assumes liability until the grain reach the port of destination chosen by the buyer. Figure 4 shows the class diagram of the standard contract INCOGRAIN-13 and semantic relationships between different actors.

4. SYSTEM ARCHITECTURE AND IMPLEMENTATION

Figure 5 shows the general architecture of the system. It includes four basic elements:

1. An RDF database (triple store), for RDF triples data storage that relate objects among them through the SPARQL query language [13]. Today, there is a list of implementations that provide the RDF triple store functionalities, including Apache Jena, Virtuoso, Owlim, Neo4J, GraphDB, Opengraph, etc. Based on the results of qualitative and quantitative study of existing RDF stores [14], we opted for the use of Apache Jena TDB [15], which is an open source framework (based on Java) for creating semantic web oriented applications.

2. A SPARQL endpoint [16], which allows applications to query information from the triple store using the SPARQL query language. The solution adopted to implement the

SPARQL endpoint is the free software Jena Fuseki [17,6]. This solution offers external applications the possibility of exploiting our data triple store by questioning directly the deposit trough SPARQL queries or even to combine it with other triple stores. Fuseki also offers two ways to interact with the user from a web application via an HTTP or with application programming interface (API). Our SPARQL endpoint provides an intuitive interface to write the SPARQL query and to select the formats of the results (JSon, XML, CSV, etc).

3. A Java Web application deployed on a WildFly application server [18] (a set of Servlets and JSP pages) has been developed to provide a simple way to view modify and enrich the data warehouse. The web interface offers two levels of navigation (conceptual level: General view of the model, instance level: details of the instances).

4. A set of transformation tools (called integrators) allowing to automatically convert data (databases, Excel files, etc.) to RDF triples (complying with the ontology previously defined).

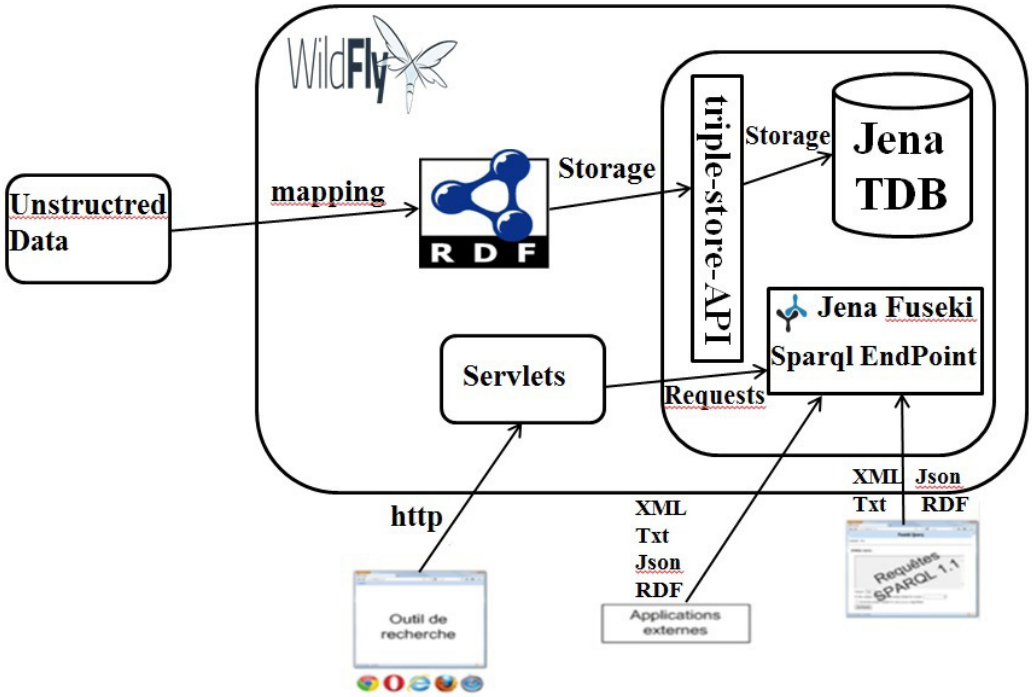


Figure 5. Detailed Architecture

5. CONCLUSIONS

This work has shown how the use of Semantic Web technologies could greatly improve the design of semantic data warehouses and facilitate the integration of data semantics by giving a formal semantics to data elements. Firstly, we have proposed a domain ontology which has given a semantic representation to all data related to the activity of grain trade. This ontology has been designed by applying the best practices proposed in [10]. Then, we have developed integrators enabling to automatically map data collected from several formats to RDF triples. Finally, a SPARQL endpoint has been proposed to provides access to the data warehouses. In the future, we plan to enrich our data warehouse with external sources available on open data platforms and to

take advantage of the power of reasoning engines from Web Semantic technologies to manage smartly the content.

ACKNOWLEDGEMENTS

This research is supported by the European Union (EU) with the European Regional Development Fund (ERDF) and Normandy Region.

REFERENCES

[1] Luca Cabibbo and Riccardo Torlone. A logical approach to multidimensional databases. In Proceedings of the 6th International Conference on Extending Database Technology: Advances in Database Technology, EDBT '98, pages 183– 197, London, UK, UK, 1998. Springer-Verlag.

[2] Aris Tsois, Nikos Karayannidis, and Timos Sellis. Mac: Conceptual data modelling for OLAP. In 3rd International Workshop on Design and Management of Data Warehouses (DMDW 2001, page 2001, 2001.

[3] Surajit Chaudhuri and Umeshwar Dayal. An overview of data warehousing and OLAP technology. SIGMOD Rec., 26(1):65–74, March 1997.

[4] Jesus Pardillo and Jose-Norberto Mazo´n. Using ontologies for the design of data warehouses. CoRR, abs/1106.0304, 2011.

[5] Khouri Selma, Boukhari Ilyès, Bellatreche Ladjel, Sardet Eric, Jean Stéphane and Baron Michael. Ontology-based structured web data warehouses for sustainable interoperability: Requirement modeling, design methodology and tool. Comput. Ind., 63(8):799–812, October 2012.

[6] Ouchenne, B. & Itmi, M. OntoEDIFACT: An Ontology for the UN/EDIFACT Standard DBKDA 2017 : The Ninth International Conference on Advances in Databases, Knowledge, and Data Applications, IARIA XPS Press, 2017, 91-96

[7] Resource Description Framework (RDF). (2016). https://www.w3.org/RDF/.

[8] Andre Menolli, H. Sofia Pinto, Sheila Reinehr, and Andreia Malucelli. An incremental and iterative process for ontology building.

[9] Mariano Fernandez, Asuncion Gomez-Perez, and Natalia Juristo. Methontology: from ontological art towards ontological engineering. In Proceedings of the AAAI97 Spring Symposium Series on Ontological Engineering, pages 33–40, Stanford, USA, March 1997.

[10] Jean Charlet, Bruno Bachimont, and Raphal Troncy. Ontologies pour le web smantique. Revue I3, page 31p, 2004.

[11] Dan Brickley and Libby Miller. FOAF Vocabulary Specification 0.99. http://xmlns.com/foaf/spec/.

[12] Christine Golbreich and Evan K. Wallace. OWL 2 Web Ontology Language New Features and Rationale. https://www.w3.org/TR/owl2-new-features/.

[13] Steve Harris and Andy Seaborne. SPARQL 1.1 Query Language. https://www. w3.org/TR/2013/REC-sparql11-query-20130321/.

[14] Bernhard Haslhofer, Elaheh Momeni Roochi, Bernhard Schandl, and Stefan Zander. European rdf store report. Technical report, University of Vienna, Vienna, March 2011.

[15] Apache Jena - TDB. https://jena.apache.org/documentation/tdb/.

[16] Lee Feigenbaum Kendall Grant Clark and Elias Torres. SPARQL Protocol for RDF. https://www.w3.org/TR/rdf-sparql-protocol/.

[17] Apache Jena Fuseki. https://jena.apache.org/documentation/fuseki2/.

[18] WildFly application server. http://www.wildfly.org/.

INTENTIONAL BLANK

DISTRIBUTED SYSTEM APPROACH TO EXPERIMENT REGIONAL COMPETITIVENESS

Mhamed Itmi[1] and Abdelkhalak El Hami[2]

[1]LITIS Laboratory, INSA Rouen Normandy,
University of Rouen Normandy. France
[2]LMN Laboratory, INSA Rouen Normandy,
University of Rouen Normandy. France

ABSTRACT

This paper highlights a work under development on a regional competitiveness project. We report on a multi-lateral, multi-scale perspective for building cooperative relationships that enhance competitiveness Regionally. The approach mimics a System of Systems methodology whereby entity relationships are captured and defined along several dimensions involving multiple constituents and multiple domain concerns. We build a serious game that is a distributed business simulator to approach the prototyping of this crossroads between supply chain management, geographical economics and information systems.

KEYWORDS

Adaptive systems, economic geography, multi agent systems, serious game, system of systems, supply chain management.

1. INTRODUCTION

In order to optimize the goods flows along the Seine River in France and to promote the territories, we need to find solutions to the current lack of quantified data but also the knowledge on the stakeholders' practices of the supply chain. In another hand a territory can be observed such as an autonomous system [1] that interacts with other territories with respect to some rules. An approach through systems of systems will be considered and used such as a serious game [2].

Some territories can differentiate themselves through a successful organization in spite of a good transport network. Logistics, thanks to infrastructure and organization, is connected to the spatial organization of the supply and distribution chain as expressed by [3]. The logistics performance is on the base of the territories competitiveness by the effect of the mutual interactions between logistics and territory. We need to prove that logistic performance boosts the territories competitiveness. To do it, we can choice the Seine axis territory because we have politicians who want to revitalize the collaboration between the economic actors of Paris, Rouen and Le Havre for the same aim: the improvement of the territory competitiveness and its sub-territories.

Our approach concerns the territorial competitiveness through the infrastructures development in particular of transport/logistics. In France and Europe, the extension of freight transportation networks will be limited in the near future due to important investments in their development

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 103–109, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80409

these last years. In socio-economics of transport, infrastructures and associated services are close. Indeed, the analysis of infrastructures allows obtaining qualitative data on freight transportation services. But we need to understand the organization around this infrastructure, organization that helps fulfil the exchange of goods.

This paper occupies a crossroads between supply chain management, geographical economics and information systems. Indeed, to respond to this problem, we need to understand why and how the complex networks notion of interaction is effective to elaborate models and simulations leading to an "intelligent" territory management. A first approach of this work appears in [2]. In the following we will discuss logistics and the economic development of territories. Then we present our approach to regional competitiveness through adaptive system of systems (SoS). We consider the economic infrastructure of a region as a global system-of-systems, economic sectors as groups and companies as systems. There are relations of dependencies between enterprises. We hint by economic infrastructure the internal facilities of a territory that ease business activity, such as communication, transportation, distribution networks and markets. We have been inspired by such approach and used a distributed serious game for the prototyping of a territory dynamics.

2. METHODOLOGICAL APPROACH

We have two approaches to study the relationship between logistics and the economic development of territories:

- An approach by territories when analysing the infrastructures that are the existing organizations,

- An approach by the interactions between all the supply chain stakeholders, all involved in a search for global territory performance.

The inland logistics is the most complex and the most important part of the whole supply chain. It is complex because supply chains develop in a moving of the spatial scales [4], in a wanted services sophistication (just in time, requirements of distribution networks), in an unstable spatial competition (off shoring and back shoring of firms) and in more complex environmental requirements.

The logistic activity is the physical and organizational continuation of the freight transport as presented by [5]. It materializes by a conurbation on a territory (agglomeration of logistic companies around a hub, idem for the information flows, etc.). The spatial agglomeration of transport and logistics activities is a reality in the harbour area such as those along the Seine axis. We shall thus develop our work on the evolution around the optimization of the logics of spatial setting-up of the logistic and distributive activities. We will analyse some types of products to understand if specificities exist as for the localization criteria of their activities. We shall focus on some types of products to analyse them in order to understand the relation customer-supplier and the impacts on the supply chain. In the Seine axis territory, we will work with manufacturers to analyse their development strategy and their relation between logistics and territory

Thanks to the analysis of the supply chain stakeholders' knowledge on a territory, we supply information to build and study systems of systems. Given the limits of calculation and engineering, the approach by the systems reliability can help to surmount these difficulties.

The architecture of the system-of-systems is modelled as a directed and operational network. The nodes represent either the component systems or a capability that needs to be acquired. Correspondingly, the links represent the dependencies of the operability between the systems or

between the capabilities. The reliability of a system-of-systems can be evaluated through the estimation of the impact of interoperability barriers in addition to the exchange inefficiency. Further analysis can be executed to assess the benefits of adding or removing systems.

3. SYSTEM ENGINEERING CONSIDERATIONS

We will deal with a multidisciplinary approach specifically suited to the context of transport and logistics.

By the very nature of the industry, a multidisciplinary approach that also considers the economic and legal dimensions of this problem is appropriate and fundamental to understand the studied phenomenon. So certain aspects of study require legal and economic reflections, and the integration of contributions from other disciplines (such as management, economy, tax system, competition, etc.) as well as understanding the legal requirements in transport, environment and customs, etc. Hence a multi-disciplinary approach will allow us to integrate several of these variable factors that impact the effective structure of such a complex system as presented in [6].

We can regard the core problem as the complex control of a complex system. The substratum, one region with a measure of autonomy, etc.) is a field where occurs an interconnected web of activities which gives rise to the production of multiple informative exchanges. It is also the place where diverse rules can be applied (economic, legal, etc.).

Diverse participants act within the framework of these rules but for their own needs and with rooms for manoeuvre. We are not studying this as a complex natural system, such as those governed by laws of the physics, and subjected to certain disturbances. In contrast, we consider a frame of reference where the entities have the freedom to act upon their own account and can be in outright conflict with other participants in the system (for resources for example).

To simplify, it operates within this context of a living system that has to remain alive reaching a certain balance. The system is in continual evolution with respect to certain rules that are changeable and hence cannot be easily modelled by a classic model. By being alive, it reacts, readjusts and modifies its hypotheses through auto-adaptation.

By such an approach, we can analyze the mass of information exchanges that we transform into knowledge. By highlighting the cognitive elements at diverse scales, we think we can give a representation of the state of the substratum with intervention onto the controllable elements

4. SYSTEM OF SYSTEMS APPROACH

We consider an adaptive system of systems and the problem of multi-scale control. The problem is to realize a real-time control according to global but multi-scale objectives of groups of heterogeneous local reactive systems varying in their behaviours, having possibility of exchange information about theirs states and behaviours to put them in a virtual self-adaptive network managing with coherence their behaviours [7]. The entire system interacting with the real world is presented in (Figure 1).

One way is to work on the reliability is to try to respond to the stimulated concerns related to reliability by assessing the infrastructure of the system-of-systems and its functional dependencies, then to evaluate the interoperability of all dependencies between interconnected systems to finally deduct their reliability. The motivation behind such approach is to inspect the structural architecture of systems-of-systems, especially the dependencies between systems in

order to evaluate, assess quantify and even anticipate (in some cases) the reliability of the dependencies within systems-of-systems.

Within the hierarchy of systems of systems we focus on an open network of proactive systems: a model with virtualization of the systems mapping the real reactive systems, with on-line control of the behaviours, links, aggregations of activities of groups of systems, management of the emergence of coalitions of actions for the global on-line pseudo-optimization of the activity of the set of proactive systems.

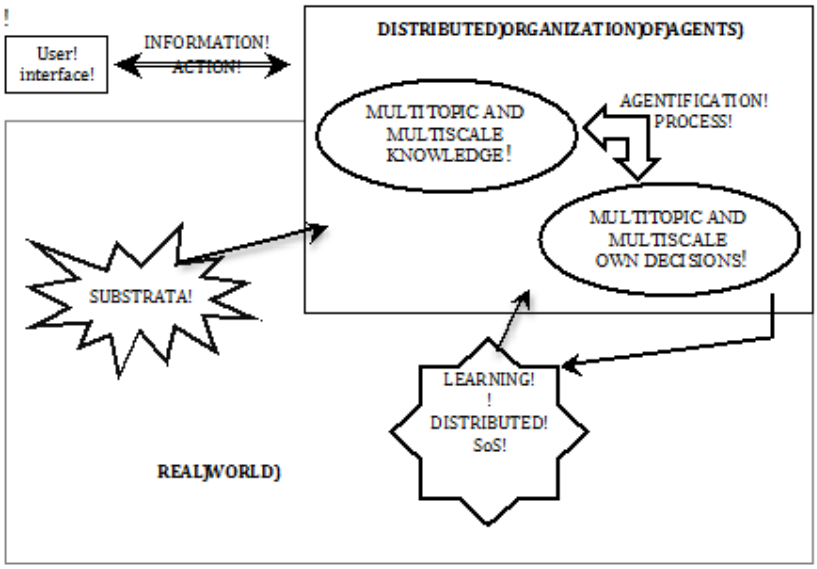


Figure 1. General architecture of the adaptive system

The basis for the construction of such a system is the agentification of the knowledge and the global goals and tendencies that allow the activity of the system in its dynamic environment. An introduction to this approach can be read in [8]. We construct an interpretation layer i.e. the agentification of both knowledge and functionalities. We determine that we know about the problems the sets of functional components have to solve. We also determine what we know about the interaction between the components at all the levels. We must use ontology for the extraction of this knowledge about states, facts and functionalities, as in classical Knowledge Based Problems: see [9]. For that, we can use the statistical analysis about the situations we have to express in the specific domain of application of the system. Then we obtain several hierarchies of knowledge and meta-knowledge with their relations.

From this first structured knowledge, we use an agentification methodology to transform the structural knowledge into a dynamic one using specific aspectual agents. In fact, we extract from knowledge all the pertinent characters of the states and relations between the system's states, and we called them semantic traits. At each semantic trait, we associate several aspectual agents expressing dynamically the pertinence of this semantic trait into the contexts of activity. We thus obtained a massive multi-agent organization of aspectual agents.

More precisely, any information in the functional system has the form of some symbolic data. We first apply a categorization about this information with transformation of information into knowledge as, for example, with the images and statistics we can use. The transformation of the basic information and physical elements behaviours into agents is not a simple one-to-one application, but an interpretation transposing symbolic structures into dynamic structures. For any

information the object system manipulates, we obtain some semantic traits expressing the characters of the knowledge this information can express in the possible contexts. So, each semantic trait is expressed with several aspectual agents. We can notice that any semantic trait has many aspectual agents matching it: the well-corresponding aspectual agents, the converse agents, the proxy agents and so on, expressing the semantic trait with a cloud made of a dynamic group of aspectual agents around the reified semantic trait.

This aspectual organization will wrap all the basic information of the object layer in order to extract its current characters. By their actions and proactivities, the active aspectual agents will generate the emergence of pertinent groups of semantic traits relative to the current behaviour and actions, taking into account the characters of their contextual relations. Each agent expresses characteristics and partial signification about the situated information contained in the active information, and the meaning of all the current information is expressed with the formation and transformation of groups of coactive aspectual agents. For the generation of this emergent agent's representation we shall use a specific kind of agents' organization management that will be a unified multi-scale control. This is a highlight of how the building work of modelling the Region activities such as an adaptive system of systems begins. The whole building of this system can be read in [1] and [10].

5. PROTOTYPING

We distinguished two approaches for the prototyping:

The first one follows the precedent description of a SoS. It needs the development of different tools introduced in the modelling (massive MAS, KBS…) and material (such as the necessary knowledge of the concerned domains). As it has been said in earlier papers, we are still working on this way.

The second approach is to go faster in the prototyping itself. In this way we relax the development constraints and develop a game demonstrator that helps the understanding of the main notions taken into account in the project. Among those notions: autonomy, information exchange and allowed context rules in the different domains are the most important items to take into account.

We went towards this last development direction. Our investigations show an interest in the subject because the work can be taken later in the education area and can be of interest for our students: first as an example of distributed simulation system. Second, as an application that can help in the understanding of the autonomy heart's mechanism. Next, we focus on the distributed simulation example tool.

The building of the tool, as a first prototype, is based on the following rules:

- To use existing simulation environments.
- Each environment should run on one PC.
- To develop on each environment an example (SME, transport company, etc.). Examples follow a business process modelling.
- To use a network supporting the simulation environments in their communication needs.

Thanks to virtualization technique one can run a first prototype in few machines. Thanks to simulation environments with blocks we can build different examples dealing with logistics, transportation, commerce, etc. that are companies. Briefly, those companies need to exchange with others (information, goods…) with respect of some rules (legal aspects). Companies can grow or decline depending on their activities, management, etc. The companies are autonomous proactive systems.

In the game one person from the group of players takes the role of network-monitor. Others ones represent different stakeholders. Each one is in charge of one company's activities. He/She can follow some overview indicators and can act by delivering "orders" (Figure 2).

Figure 2. The serious game

The network-monitor is communicating on the global system. He/She is in charge of global rules to be respected. Playing with those rules can modify the state of the eco-system. One objective of the monitor is to maintain or improve the activity of the global system which means to respect some global indicators without loss of balance. That is usually the role of a political decision maker. Then we can observe the different dynamics and particularly the behaviour of the whole system and go towards the study of the reliability of systems-of-systems and its relationship with systems interoperability which is a relatively newly emerging field of research.

6. CONCLUSIONS

Thanks to systems of systems approach based on the agentification and knowledge, we plan to represent the interconnected management and decision entities of a Region and give a frame to the competitiveness notion. A serious game can help to better understand the territory dynamics and also to go further in the understanding of systems-of-systems reliability and interoperability, and particularly resilience quantification.

ACKNOWLEDGEMENTS

This research is supported by the European Union (EU) with the European Regional Development Fund (ERDF) and Normandy Region.

REFERENCES

[1] Cardon, A. & Itmi, M. New Autonomous Systems John Wiley & Sons, 2016

[2] Verny J., Itmi M., El Hami A., Cardon A., Couturier L. and Abdulrab H., "A Sustainable multidisciplinary Approach to Building Regional Competitiveness". In the Symposium on security and safety of Complex Systems, 2SCS'12. 25-26 May. Agadir, Morocco. 5p. (2012)

[3] Joignaux, G., (2008). "Quel impact logistique sur l'aménagement territorial ?", Notes(Billets) de synthèse du SESP, 168. Pp. 45-50.

[4] Verny, J. and Grigentin, C., (2009). "Container shipping on the Northern Sea Route", International Journal of Production Economics, 122 (1), 107-117.

[5] Wackermann, G. (dir)., Verny, J. et Al-., (2011). Environnement et écosociété (dictionnaire), Ellipses, Paris.

[6] Axelrod, R. (1997). The complexity of Cooperation: Agent-based Model of Competition and Cooperation. Princeton: University Press.

[7] Keating, C., Rabadi, G., Landaeta, R. E., Bowling, S. (2009). "System of systems engineering for border security and immigration: methodologies, processes and tools". International Journal of System of Systems Engineering (IJSSE), Volume 1 - Issue 4.

[8] Wooldridge, M. and Jennings, N. (1995) "Intelligent Agents: Theory and Practice", Knowledge Ang. Rev., vol. 10, no.2, pp.115-152.

[9] Lenat D. and Guha R., (1990). Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project, Addison-Wesley Publishing.

[10] Itmi, M. and Cardon A. (2012). "Autonomy and Control of Adaptive Systems of Systems", International Journal of Modeling, Simulation and Scientific Computing 3(1): 1240002. 21 pages. World Scientific Publishing Company. DOI: 10.1142/S1793962312400028.

AUTHORS

Mhamed Itmi earned his PhD in Probability Theory and Statistics in 1980 and second PhD in Computer Science in 1989. He received his Habilitation Diploma to supervise research (HDR) in 2006 with the focus on the modelling and simulation of distributed discrete event systems. He managed different logistics and transportation research projects and supervised several PhD theses. He also is the author and co-author of more than 100 papers published in international journals, conferences and books. His research presently focuses on autonomous systems. He is an Associate Professor at the INSA-Rouen, France.

Abdelkhalak El Hami is a Full Professor at INSA Rouen, Normandy France, as well as Deputy Director of LMN and director of mechanical engineers. He's research activities include reliability-optimization systems. He has supervised 38 PhD theses. He also is the author and co-author of more than a twenty books and more than 550 papers published in international journals and conferences. He has a doctorate in engineering sciences from the University of Franche-Comté in France (1992). He received his Habilitation diploma to supervise research (HDR) in 2000. He's Editor in chef of 3 Set of international Book, ISTE, Wiley and Elsivier.

INTENTIONAL BLANK

RELIABILITY OF MECHANICAL SYSTEM OF SYSTEMS

El Hami Abdelkhalakl and ITMI Mhamed

LMN-INSA Rouen Normandie, Normandy University,
ST Etienne du Rouvray FRANCE

ABSTRACT

In this paper, we present a new methodology about reliability of systems of systems. We present also an example which combines the information transformation in complex systems and virtual design of this system based on finite element analysis. This example is help to balance the performances and the costs in complex system, or provide the optimal solution in manufacturing design. It can also update the existing design of component by changing the new design of this component.

KEYWORDS

Reliability, Systems of systems, Simulation

1. INTRODUCTION

The manufacturing industrials of car or aircraft are the term that covers a wide range of companies and organizations involved in the design, development, manufacture, marketing, and selling. To minimize the manufacturing costs of the relative products, some virtual methods are developed to numerically investigate the product manufacturing process through all the fields mentioned above. This report tries to describe the information transformations between different subsystems in the complex manufacturing system and introduces the finite element simulation in the numerical predictions of some mechanical structures or processes. Meanwhile, the differences between system of systems and the complex mechanical system are briefly discussed.

2. COMPLEX MECHANICAL SYSTEMS IN AIRCRAFT MANUFACTURING INDUSTRIALS

The aerospace manufacturer is a high technology industry which is a company or individual involved in the various aspects of designing, building, testing, selling, and maintaining aircraft, aircraft parts, missiles, rockets or spacecraft. Aircraft manufacturing is one of aerospace manufacturing and it is important in civil, industry, military and scientific research. This report focuses on introducing the design of global aircraft mechanic system according to various sub mechanical systems. The finite element simulation is proposed to numerically study the properties of each sub mechanical systems and virtually research the global properties of aircraft

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 111–120, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80410

system (client requirements: cost and strength for example); A platform of system is needed to transfer the information from requirements to the end manufacturing.

Aircraft manufacturing system is a complex system which includes various sub systems. The performance of this global system is closely related to each subsystem. For instance, the mechanical properties of airplane are depended on the mechanical properties of each subsystem. Figure 1 shows us the strength design of global airplane and its components. In this report, we try to integrate the strength design of subsystems or components which predicted by the finite element simulations into the global strength in order to fulfill the different needs of customers. All of the resources like manufacturing machines, metal forming processes, materials, and customer requirements should be integrated together in this virtual tool.

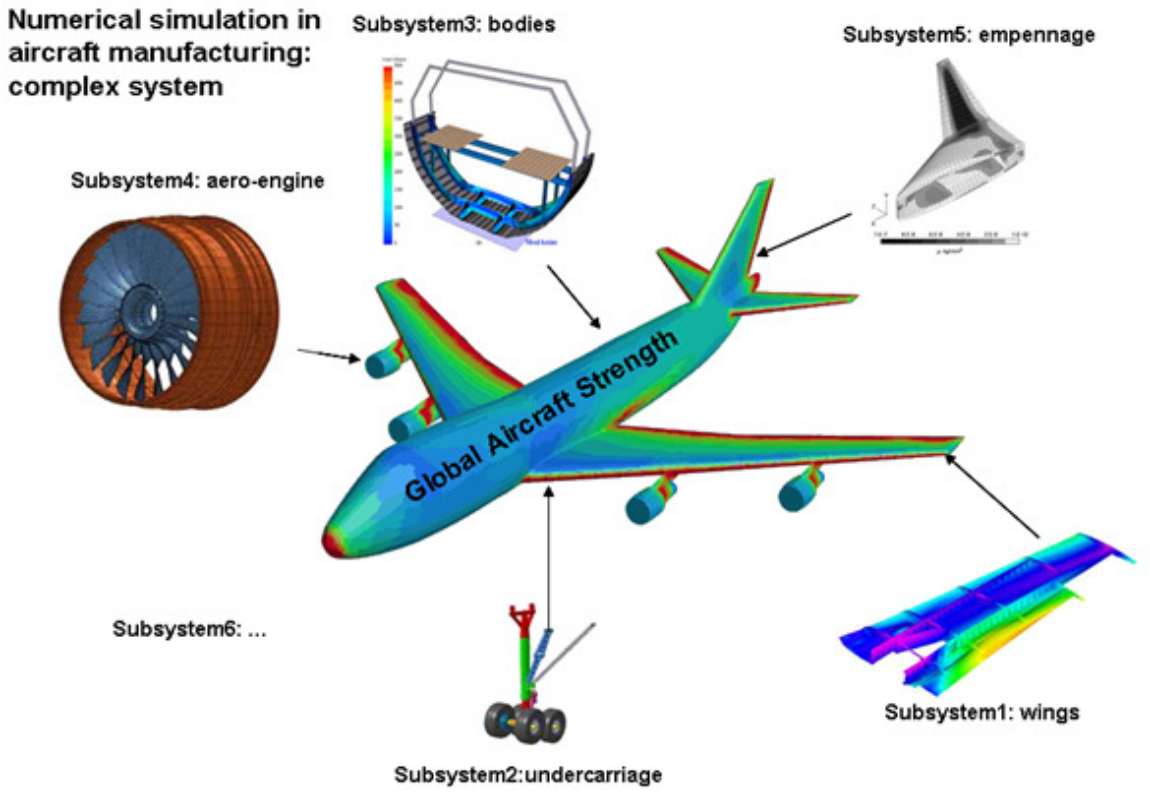


Figure 1. Global aircraft strength and the strengths in different parts.

In the global strength design of aircraft, the strength, the weight and the costs are three important aspects must be taken into account. Generally speaking, the cost should be controlled in a certain range, while the total weight is as light as possible and the strength must fulfill the demand of customers during the aircraft design process. This will demand firstly to transfer the global information into each subsystems or components and then focus on the component productions, like undercarriage, wings, aero-engines and empennage. Each component will have its own demand on strength, weight and costs. This is the aspect of the assignment of customer requirements from global strength design to component strength design. The other aspect is to numerically predict the strength of component with different design schemes and chose the optimal one based on the finite element simulation and other optimization algorithms. For

example in wing design (see Figure 2), the optimal design scheme can be obtained with the help of finite element simulation. Some performances like, loading capacity, fatigue performance and aerodynamic structure can be numerically optimized for industrials.

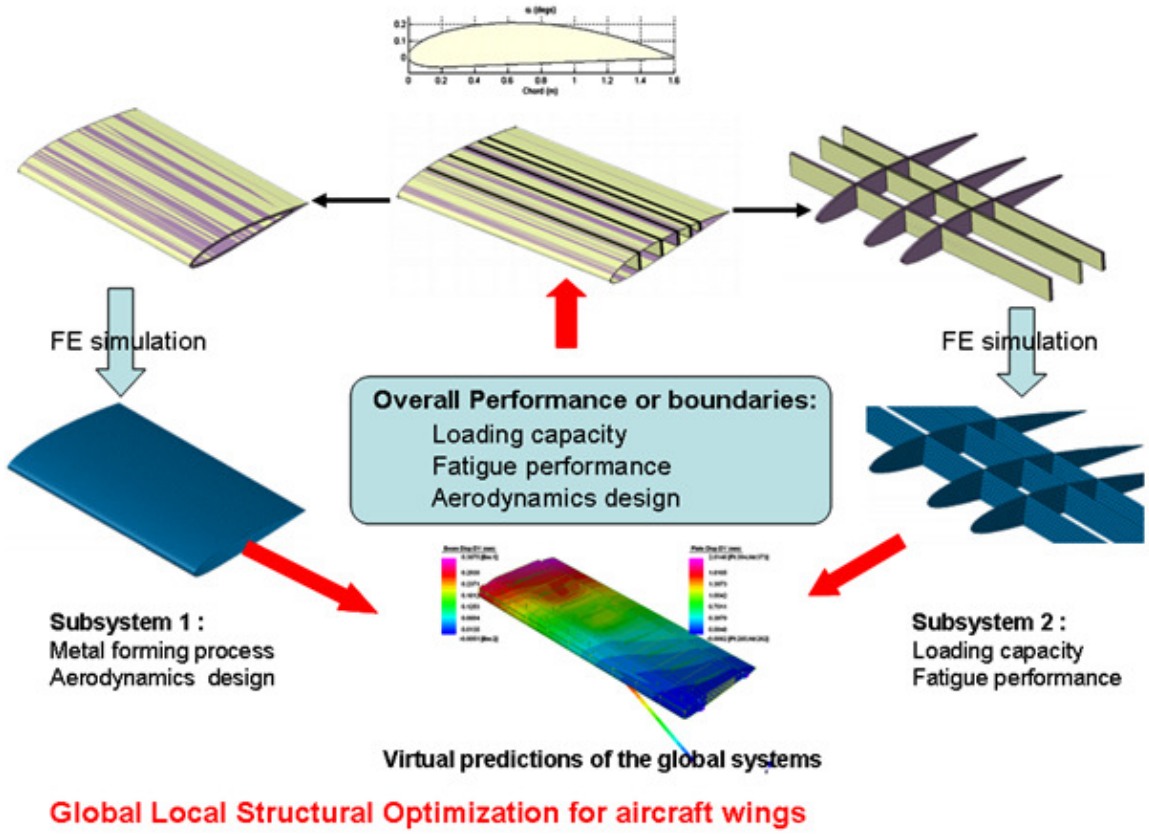


Figure 2. Wing's design and its strength simulation with the help of finite element simulation.

Structure design:

1. Aerodynamic design;
2. Structure strength;

Materials:

1. Strength-to-weight ratio ;
2. Cost: metals, composite material, new function material…;
3. Elastoplastic-damage response in complex conditions: temperature, fatigue…;

Metal forming process

1. Metal forming and piece strength;
2. Machining cost
3. Machining or forming time: design cycle…;

Rivet connection:

1. Connection strength: loading, fatigue…;
2. Stress concentration;
3. And so on….

3. SIMULATIONS

The aspects listed above (structure design, material chosen, forming processes, rivet connection…) can be studied in the finite element simulation and their mechanical properties can be numerically predicted. Some numerical simulations of components which can be used as the indicators are shown in Figure 3 (rivet connection), Figure 4 (strength of thread) and Figure 5 (pipe bending process).

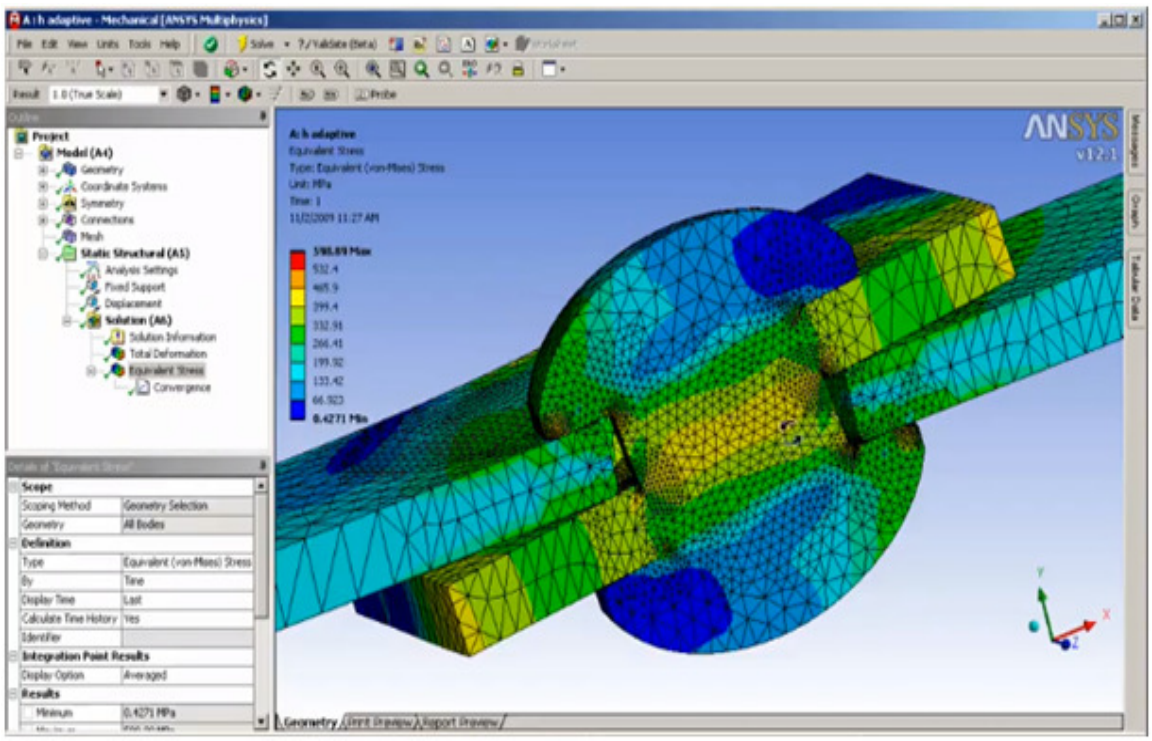


Figure 3. The prediction of linking strength through Finite Element Simulation

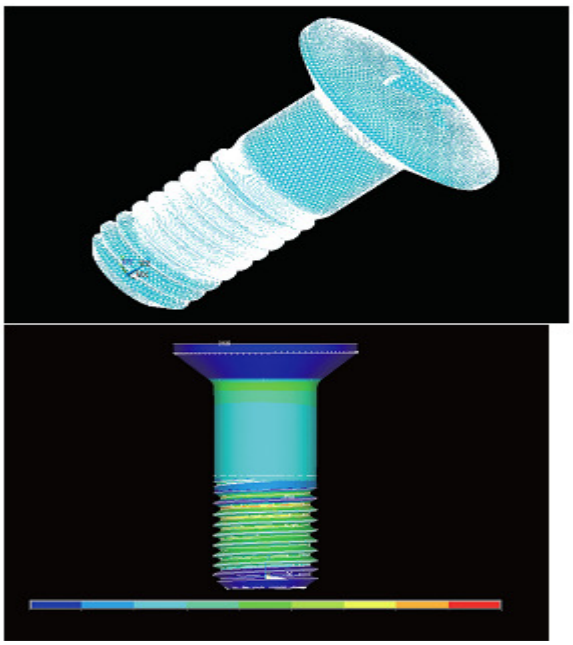


Figure 4. The prediction of stress concentration in thread through Finite Element Simulation

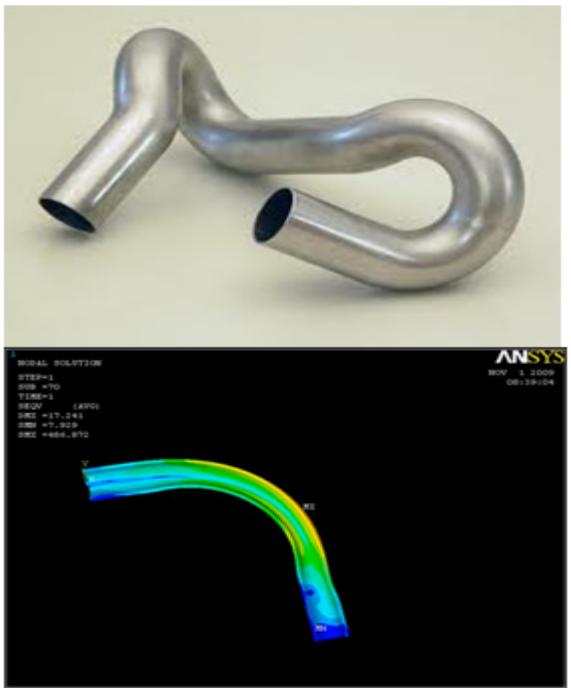


Figure 5. The prediction of pip bending process for aero-engine through Finite Element Simulation

Based on the numerical results about the components of aircraft, the global strength, weight and cost will be obtained when these components are assembled together, as shown in Figure 6.

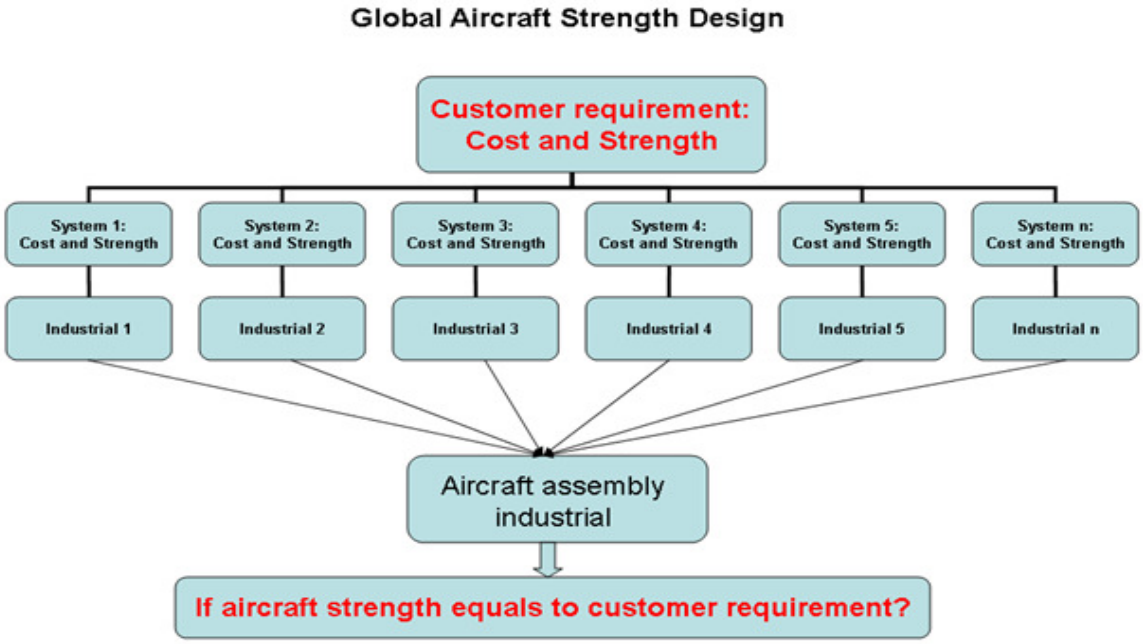


Figure 6. Assembly of the subsystems

The numerical predictions based on finite element simulation for each component can virtually study the manufacturing and designing processes of components. A platform of system is also needed to transfer requirements (cost, weight and strength) to from global design to the end manufacturing, as follows:

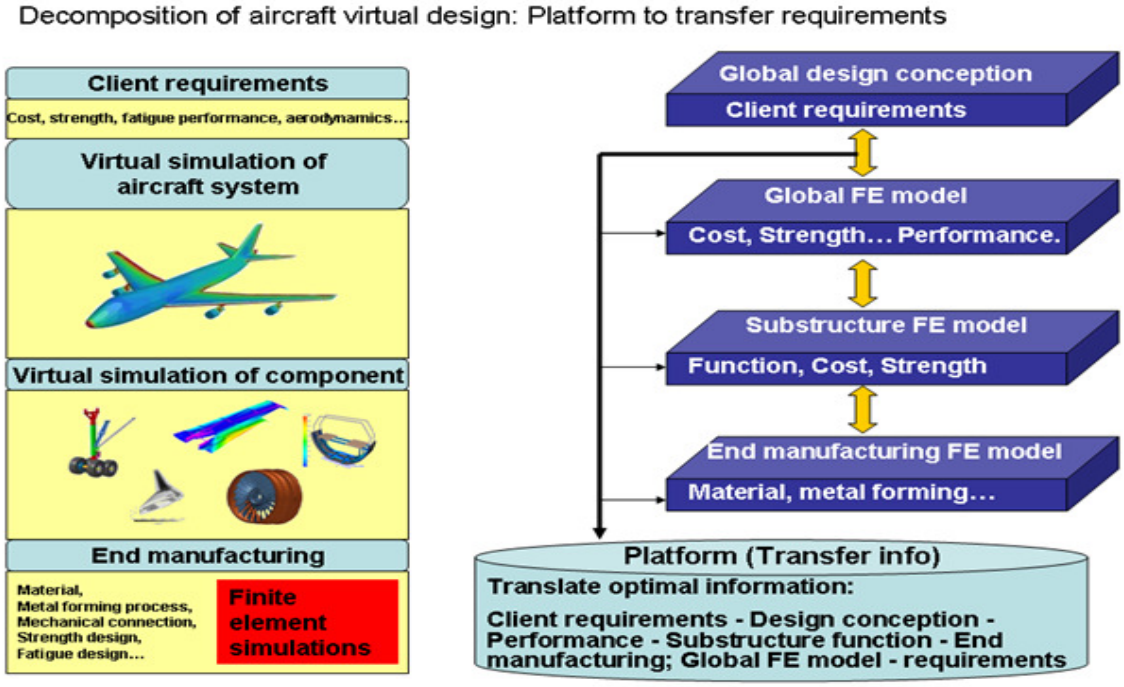


Figure 7. Information transformation between global requirements and the requirements in subsystems

In this report, we try to present an example which combines the information transformation in complex systems and virtual design of this system based on finite element analysis. This example is help to balance the performances and the costs in complex system, or provide the optimal solution in manufacturing design. It can also update the existing design of component by changing the new design of this component.

4. NUMERICAL PLATFORMS BASED ON FINITE ELEMENT SIMULATION

4.1 Numerical Platform

Based on the above description, a numerical platform based on finite element simulation must be build in order to numerically study the mechanical systems. It will be used firstly to simulate metal forming processes and consider the user defined material subroutine which implements advanced material constitutive equations. Secondly, it can be used to simulate the responses (strength, efficiency, fatigue, properties…) of the mechanical structures or connections…. The implementation of this platform is based on ANSYS platform and is described in the following flowchart.

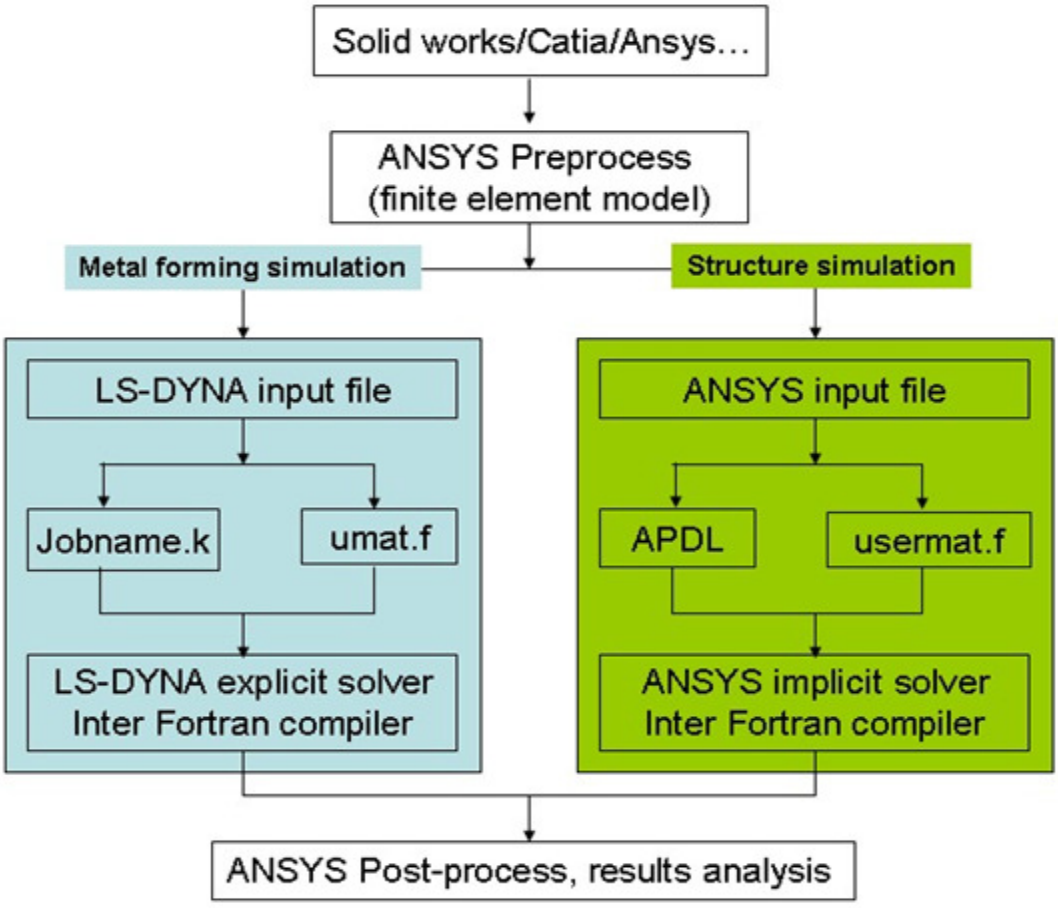


Figure 8. Finite element simulation platform based on Ansys

5. CONCLUSIONS

System of systems in the information field is a collection of task-oriented or dedicated systems that pool their resources and capabilities together to create a new, more complex system which offers more functionality and performance than simply the sum of the constituent systems. Comparing to system of systems, the complex mechanical system is a product (car, airplane and so on) which is composed of some components. Unlike the system of systems, the complex mechanical system cannot leave any components, in other words, each component has its function and their functions are different. The virtual simulation of each component based on finite element method is surly useful in the design and manufacture of complex mechanical system, but it is difficult to used in system of systems according to its conception.

Actually, the numerical simulation is a suitable tool to do the reliability analysis for the mechanical systems. From this point of view, the numerical predictions can be used in system of systems as an indicator of the reliability of subsystems. The finite element analysis can be used to estimate the abilities (strength, fatigue, damage) of each subsystem. For example in the ballistic missile defense system, finite element simulation can be used to predict the mechanical reliability of each defense weapon, or to estimate the usability of this subsystem in the mechanical aspect. However, in the system of systems, all of the subsystems are assumed to be usable and the research key point is its functions. Hence, the bridge between the field of system of systems and the filed of mechanical systems is the requirement of the reliability analysis of mechanical system in the system (of systems).

ACKNOWLEDGEMENTS

This research is supported by the European Union (EU) with the European Regional Development Fund (ERDF) and Normandy Region.

REFERENCES

[1] Y. Bao, T. Wierzbicki, On fracture locus in the equivalent strain and stress triaxiality space, International Journal of Mechanical Sciences, 46 (2004) 81-98.

[2] Y. Bao, T. Wierzbicki, On the cut-off value of negative triaxiality for fracture, Engineering Fracture Mechanics, 72 (2005) 1049-1069.

[3] T. Wierzbicki, Y. Bao, Y.-W. Lee, Y. Bai, Calibration and evaluation of seven fracture models, International Journal of Mechanical Sciences, 47 (2005) 719-743.

[4] Y. Bai, Effect of Loading History on Necking and Fracture, in, Massachusetts Institute of Technology, 2008.

[5] Y. Bai, T. Wierzbicki, A new model of metal plasticity and fracture with pressure and Lode dependence, International Journal of Plasticity, 24 (2008) 1071-1096.

[6] D.C.a.P. Drucker, W., Soil mechanics and plastic analysis for limit design., Quarterly of Applied Mathematics, 10 (1952) 157-165.

[7] F.A. McClintock, A Criterion for Ductile Fracture by the Growth of Holes, Journal of Applied Mechanics, 35 (1968) 363-371.

[8] J.R. Rice, D.M. Tracey, On the ductile enlargement of voids in triaxial stress fields, Journal of the Mechanics and Physics of Solids, 17 (1969) 201-217.

[9] A.L. Gurson, Continuum Theory of Ductile Rupture by Void Nucleation and Growth: Part I---Yield Criteria and Flow Rules for Porous Ductile Media, Journal of Engineering Materials and Technology, 99 (1977) 2-15.

[10] J.L. Chaboche, Anisotropic creep damage in the framework of continuum damage mechanics, Nuclear Engineering and Design, 79 (1984) 309-319.

[11] J. Lemaitre, Coupled elasto-plasticity and damage constitutive equations, Computer Methods in Applied Mechanics and Engineering, 51 (1985) 31-49.

[12] J.C. Simo, J.W. Ju, Strain- and stress-based continuum damage models—I. Formulation, International Journal of Solids and Structures, 23 (1987) 821-840.

[13] K. Saanouni, On the numerical prediction of the ductile fracture in metal forming, Engineering Fracture Mechanics, 75 (2008) 3545-3559.

[14] J. Lemaitre, A course on Damage Mechanics, Springer-Verlag, New York, 1996.

[15] W.A. Spitzig, O. Richmond, The effect of pressure on the flow stress of metals, Acta Metallurgica, 32 (1984) 457-463.

[16] J.C. Simo, R.L. Taylor, Consistent tangent operators for rate-independent elastoplasticity, Computer Methods in Applied Mechanics and Engineering, 48 (1985) 101-118.

[17] J.C. Simo, R.L. Taylor, K.S. Pister, Variational and projection methods for the volume constraint in finite deformation elasto-plasticity, Computer Methods in Applied Mechanics and Engineering, 51 (1985) 177-208.

[18] J.L. Chaboche, G. Cailletaud, Integration methods for complex plastic constitutive equations, Computer Methods in Applied Mechanics and Engineering, 133 (1996) 125-155

AUTHORS

Abdelkhalak El Hami is a Full Professor at INSA Rouen, Normandy France, as well as Deputy Director of LMN and director of mechanical engineers. He's research activities include reliability-optimization systems. He has supervised 38 PhD theses. He also is the author and co-author of more than a twenty books and more than 550 papers published in international journals and conferences. He has a doctorate in engineering sciences from the University of Franche-Comté in France (1992). He received his Habilitation diploma to supervise research (HDR) in 2000. He's Editor in chef of 3 Set of international Book, ISTE, Wiley and Elsivier.

Mhamed Itmi earned his PhD in Probability Theory and Statistics in 1980 and second PhD in Computer Science in 1989. He received his Habilitation Diploma to supervise research (HDR) in 2006 with the focus on the modelling and simulation of distributed discrete event systems. He managed different logistics and transportation research projects and supervised several PhD theses. He also is the author and co-author of more than 100 papers published in international journals, conferences and books. His research presently focuses on autonomous systems. He is an Associate Professor at the INSARouen, France.

AUTHOR INDEX

Abdelkhalak El Hami 103

Boulares Ouchenne 93

Devesh C. Jinwala 53

El Hami Abdelkhalakl 111

Farhad Foroughi 01

Ghada Badr 15

ITMI Mhamed 111

Kaushal Shah 53

Khalid Alsamara 37

Manish Kumar 69

Mhamed Itmi 93, 103

Noura A. Semary 81

Peter Luksch 01

Ruchika Gupta 69

Saleem Abuleil 37

Sara Al-Osimi 15

Udai Pratap Rao 69

Yong-Jae Kim 47