OCTA

6-8 February, 2020

ALECSO, Tunis -Tunisia

OCTA International Multi-Conference on:

Organization of Knowledge and Advanced Technologies

https://multiconference-octa.loria.fr/

Unifying the scientific contributions of the following conferences:

SIIE

8th. International Conference on:

Information Systems and Economic Intelligence

ISKO-Maghreb

Digital Sciences: impacts and challenges on Knowledge Organization

CITED

2nd. International Symposium on:

Advanced Technologies, Renewable Energies and Economic Development

TBMS

1st. international symposium on:

Big-Data-Analytics Technologies for Strategic Management: innovation and competitiveness

ORGANIZERS

Université de Tunis

ISKO-Maghreb

UNIVERSITÉ DE LORRAINE

IEEE Tunisia Section

www.ieee.tn

المنظمة العربية للتربية والثقافة والعلوم

SPONSORS

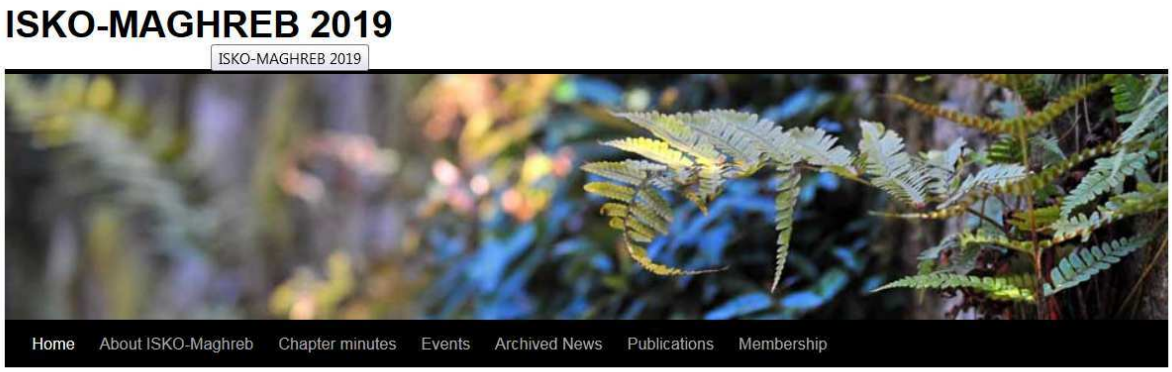
Tunisia China Friendship Association

ATLAS

SPECIAL ISSUES JOURNALS & BOOKS

isprs International Journal of Geo-Information

IEEE Xplore® Digital Library



PROCEEDINGS

8th. Edition of ISKO-Maghreb'2019 on: "Digital Sciences: impacts and challenges on Knowledge Organization".

Organization in the event of the International Multi-Conference OCTA'2019 on: « Organization of Knowledge and Advanced Technologies » (https://multiconference-octa.loria.fr/)

February 6-7-8, 2020 Tunis (Tunisia)

www.isko-maghreb.org **&** https://isko-maghreb2019.loria.fr

ISKO-Maghreb2019 CFP & Special issues in Journals & Books

Description :

"*The world is changing*" according to **Michel Serres** in his book "Petite Poucette" (March 2012) and in his detailed speech at the French Academy (November 2017), he argues by: "*It should then draw the consequences of this change of the space that affects the human species with the emergence of the "new technologies" of information and communication – in particular by trying to establish a new balance between the material element and the intangible element. "*. Historically for Humanity, two major inventions that have profoundly changed the **knowledge organization** (KO) while associating culture and the transmission of knowledge: it is about writing (around the fourth millennium BC), then printing (in the fifteenth century by the invention of Johannes Gutenberg). Today, with digital technology, it is a major new invention that is imposing and having profound repercussions on the **knowledge organization**

system (KOS) and its transmission of knowledge, on teaching and pedagogy, and on society and its world economy oriented towards the intangible.

At the dawn of the twenty-first century, we are witnessing the emergence of a young science called "digital sciences" to designate the information and communication sciences on their hardware and software components. The digital sciences revolutionize knowledge in relation to its technological means, as well as social links (or humanity) with its relationship to knowledge and its organization.

To the changes brought about by digital sciences, society as a whole must adapt and consider skills that reinvent tomorrow's social relationships and its relation to knowledge: what we call the dynamic and interactive Web (web2.0 and web3.0), the Social networks that allow the creation of educational, professional and societal communities, the Collective Production of Knowledge and its relation to the collective intelligence, the risks of the Proliferation of the information and its consequences like the misinformation, the "fake news" or invasion of privacy. All of these changes have a potential and profound impact on education, culture, society and knowledge organization systems (KOS).

Faced with the advent of the digital sciences, the governance of knowledge seems to be the scientific policy best suited to the creation of value with regard to Man and his evolution in cultures and civilizations. The task of this governance is to take into consideration the transmission of knowledge related to scientific, technological and communication progress. Intrinsically, this process of knowledge transfer requires a system of knowledge organization and management that implements the production of knowledge, its actors and the digital sciences for its influence in society.

The objective of the ISKO-Maghreb chapter continues to contribute to understanding the factors that organize knowledge and the phenomena that affect the information society. The actions to be undertaken by the scholarly society ISKO will have to take into account socio-cultural, cognitive and economic aspects in the strategic management of knowledge. Towards the knowledge society, knowledge must be seen in its dynamics, content and scientific and technological interactions with academics, business and politics (actors and institutions).

In this context, a first orientation is pedagogical to try to answer the question "***what do we know about knowledge, its organization and its mutation***?". Then, the question evolves towards the societal challenges of knowledge, in theory and in practice, to provide clarifications to "***what convergence of KO (Knowledge Organization) and KM (Knowledge Management) approaches that organize knowledge and know-how?*** ". Education, Digital Sciences, Culture, Information and Communication Sciences, remain the major themes covered by ISKO-Maghreb, for the establishment of knowledge organization, management skills, collective intelligence and Digital Humanities.

In a friendly, warm and open to exchange, the ISKO-Maghreb learned society was designed to strengthen the "ISKO International" Scholarship Foundation with academics, practitioners both in the Maghreb countries and in the world.

Committees

Steering Committee OCTA'2019 :

Honorary President
Pr. Habib SIDHOM, President of the University of Tunis (Tunisia)

Presidents
Sahbi SIDHOM (LORIA – Université de Lorraine, France)
Amira KADDOUR (ENSTAB – Université de Carthage, Tunisia)
Anass EL HADDADI (ENSA Al-Hoceima, Morocco)
Mohamed ADDAM (ENSAH Al Hoceima, Morocco).
Abdelkrim MEZIANE (Centre de Recherche CERIST, Algeria)
Davy MONTICOLO (ENSGSI – Université de Lorraine, France)
Xi LU (Tsinghu University, China & Harvard University, Cambridge MA, USA)
Saoussen KRICHEN (Vice-President Université de Tunis, Tunisia)
Khaled KCHIR (Vice-President Université de Tunis, Tunisia)

Program committee / Comité scientifique :

Presidents

SIDHOM Sahbi (President ISKO-Maghreb, Université de Lorraine, France)
MEZIANE Abdelkrim (Vice-President ISKO-Maghreb, CERIST, Algeria)
EL HADDADI Anass (Vice-President ISKO-Maghreb, ENSA Al-Hoceima, Morocco)
KADDOUR Amira (Vice-President ISKO-Maghreb, Université de Carthage, Tunisia)

Members

ADDAM Mohamed (ENSA, Al-Hoceima, Maroc)
AHMED-ZAID-CHERTOUK Malika (REDYL Lab. – Université de TIZI-OUZOU, ALGERIE)
AHROR Belaid, (D.Informatique Université de Bejaia, Algeria)
AJHOUN Rachida (ENSIAS, Université Mohammed V Souissi, Rabat, Maroc)
ALIANE Hassina (CERIST, Algeria)
ALOUI Abdelouahab, (Université de Bejaia, Algeria)
AMROUN Kamal, (Université de Bejaia, Algeria)
AOUGHLIS Farida (Université M.Mammeri-Tizi Ouzou, Algeria)
AYAD Soheyb, (LINFI Lab., Université de Biskra, Algeria)
AZNI Mohamed, (Université de Bejaia, Algeria)
BABIK Wieslaw (Jagiellonian University of Cracow, Poland)
BADACHE Nadjib (Directeur du CERIST, Algeria)

BELLAMINE Narjes (Lab. RIADI – ENSI Université de la Manouba, Tunisie)
BEN AHMED Mohamed (FST, Tanger, Maroc)
BEN GHEZALA Henda (Lab. RIADI – ENSI Université de la Manouba, Tunisie)
BEN ROMDHANE Mohamed (ISD – Université de la Manouba, Tunisie)
BENABOU DJILALI (Université de Mascara, Algérie)
BENALI Khalid (Université de Lorraine, Nancy, France)
BENATALLAH Boualem, (UNSW, Australie)
BENBRAHIM Houda (ENSIAS, Université Mohammed V Souissi, Rabat, Maroc)
BENMOHAMMED Mohammed, (Université de Canstantine 2, Algeria)
BENNOUI Hammadi, (LINFI Lab., Université de Biskra, Algeria)
BERRADA Ilham (ENSIAS – Université Mohammed V Souissi, Rabat, Maroc)
BESSAI Fatma Zohra (CERIST, Algérie)
BOUDHIR Anouar Abdelhakim (FST, Tanger, Maroc)
BOUDJLIDA Nacer (Université de Lorraine, Nancy, France)
BOUHORMA Mohammed (FST, Tanger, Maroc)
BOUKERRAM Abdellah, (Université de Bejaia, Algeria)
BOUREKKACHE Samir (University of Mohamed Khider, Biskra, Algeria)
BOUSSAID Omar (Laboratoire ERIC, Université Lumière Lyon2, France)
BOYER Anne (LORIA – Université de Lorraine, France)
BRUN Armelle (LORIA – Université de Lorraine, France)
CHAIB Bachir (University 20th august 1955 Skikda, Algeria)
CHAROY François (Université de Lorraine, Nancy, France)
CHEBBI Aïda (ISD – Université de la Manouba, Tunisie)
CHEIKHI Laïla (ENSIAS, Université Mohammed V Souissi, Rabat, Maroc)
DJERAD Nejoua (ISD Université de Manouba, Tunisia)
EL BOUHDIDI Jaber (ENSA, Tétouan, Maroc)
EL BOUHISSI Houda (Université de Bejaia, Algeria)
EL GHOUL Mansour (Université de Lorraine, France)
EL HACHANI Mabrouka (ELICO – Université Jean Moulin Lyon3, France)
ESSAAIDI Mohammad, (ENSIAS, Université Mohammed V Souissi, Rabat, Maroc)
FARAH Zoubeyr, (Université de Bejaia, Algeria)
FEKI Jamel (Faculté des Sciences Economiques et de Gestion à Sfax, Tunisie)
FENNAN Abdelhadi (FST, Tanger, Maroc)
FENNICHE Raja (ISD Université de Manouba, Tunisia)
FRISCH Muriel (ESPE/Université de Reims Champagne-Ardenne, France)
GDOURA Wahid (ISD Université de Manouba, Tunisia)
GNOLI Claudio (University of Pavia, Italy)
GODART Claude, (Université de Lorraine, Nancy, France)
GRIVEL Luc (Université de Paris I Panthéon-Sorbonne, France)
HABACHA Anja (Lab. RIADI – ENSI Université de la Manouba, Tunisie)
HATON Jean-Paul (Université de Lorraine et Institut Universitaire de France, France)
HEYOUNI Mohammed (ENSA, Oujda, Maroc)
JALLOULI Rim (ESEN – Université de la Manouba, Tunisie)
JAMOUSSI Yacine (Lab. RIADI – ENSI Université de la Manouba, Tunisie)
JBENIENI Jihene (Université de Carthage, Tunisie)
KASSOU Ismaïl (ENSIAS – Université Mohammed V Souissi, Rabat, Maroc)
KAZAR Okba (Université de Biskra, Algérie)
KHALIDI IDRISSI Mohamed (Université Mohammed V Agdal, Rabat, Maroc)
KHANOUCHE Mohamed Essaid, (Université de Bejaia, Algeria)
LAGHROUCH Mourad (LAMPA – Université M.Mammeri-Tizi Ouzou, Algérie)
LAMBERT Philippe (Institut Jean Lamour – Université Nancy1, France)

LAMIREL Jean-Charles (INRIA & Université de Strasbourg, France)
LIQUETE Vincent (University of Bordeaux, France)
LÓPEZ-HUERTAS María (Vice-President of ISKO Int. & Universidad de Granada, Spain)
MAAMAR Zakaria (Zayed University, EAU)
MAHMOUD Seloua (ISD Université de Manouba, Tunisia)
MAZZOCCHI Fulvio (Institute for Complex Systems Roma, Italy)
MEDJAHED Brahim, (Michigan University, USA)
MONTICOLO Davy (ENSGSI-ERPI, Université de Lorraine, France)
NACER Hassina (Université des sciences et de technologie USTHB Alger, Algeria)
NOURRISSIER Patrick (NetLorConcept, France)
OHLY Peter (ISKO Int. & GESIS – Leibniz-Institut Bonn, Germany)
ORRICO Evelyn (UNESP-São Paulo State University, Brasil)
OUBRICH Mourad (President of CIEMS – Rabat, Maroc)
OUERGHI Feriel (Université de Tunis, Tunisie)
OUERHANI Salah (Université de Tunis, Tunisie)
Oulad Haj Thami Rachid (ENSIAS, Université Mohammed V Souissi, Rabat, Maroc)
PECQUET Pascal (University of Montpellier, France)
RABHI Fethi (UNSW, Australie)
RAMDANI Mohamed (Université Hassan II Mohammedia, Maroc)
REZEG Khaled, (LINFI Lab., Université de Biskra, Algeria)
Rodríguez-Bravo Blanca (University of León, Spain)
ROUDIES Ounsa (Ecole Mohammadia d'Ingénieurs, Université Mohammed V Agdal, Rabat, Maroc)
SALEH Imad (Université Paris 8, France)
SAN SEGUNDO Rosa (University of Madrid Carlos III, Spain)
SAYEB Yemna (ENSI – University of Manouba, Tunisia)
SEBAA Abderrazak, (Université de Bejaia, Algeria)
SEGHIR Yousra (ISD – Université de la Manouba, Tunisie)
SLIMANI Hachem, (D.Informatique Université de Bejaia, Algeria)
SOUZA Renato (UNESP-São Paulo State University, Brasil)
SYLBERZTEIN Max (LASELDI – Université de Franche-Comté, France)
TANTI Marc (Centre d'épidémiologie des armées – SESSTIM, France)
TARI Abdelkamel, (Université de Bejaia, Algeria)
TIETSE Samuel (MICA, Université François Rabelais de Tours, France)
ZELLOU Ahmed (ENSIAS, Université Mohammed V Souissi, Rabat, Maroc)

Social networking application, visual communication system for the protection of personal information

Marilou Kordahi

Abstract— We contribute to the field of Information Systems by attempting to develop a first innovative social networking application, the “SignaComm”. The SignaComm's objective is to enable multilingual communication between users worldwide for the protection of personal data on the Web. We design this application while relying on the theory of patterns as well as the principles of ontologies and signage system. The theory of patterns presents good practices for creating a model, which describes the characteristics of a generic solution to a specific problem. It permits the reuse and remodelling of patterns to serve as resources for software development and problem solving. Ontologies describe a structured set of concepts and objects by giving a meaning to an information system in a specific area, and allow the construction of relationships between these concepts and objects. The signage system is a visual communication system with an international vocation where, the “signagram” is the writing unit. When creating the SignaComm we use an automatic translation software of key-phrases into signagrams. The social networking application is written with the PHP and Javascript programming languages and then tested technically. We hope that users from any culture, social environment or with disabilities could use it.

Index Terms— Communications applications, picture/image generation, information resource management, information systems, social networking

1 Introduction

In this paper, we contribute to the field of Information Systems by analysing the connections between Social Networking Sites (SNS) and artificial communication systems (e.g., visual communication systems). Several Web applications, namely Social Networking Sites and virtual communities, are currently using artificial visual communication systems to facilitate interactions and knowledge exchange between different users and members worldwide. This may be justified by the emergence of global social developments [1] as well as an available international audience. For example, one can notice the presence of visual communication systems in the Web applications' user-interface and dictionary of emoticons [2].

In the following paragraphs we will introduce the SNS as well as artificial communication systems.

Social Networking Sites: An SNS is an online information system for building social relationships between individuals (or organisations) sharing interests, activities, contacts in real life [4], [5], [6], [7], [8], [9]. The growth of these social ties (strong or weak ties) can only take place if these individuals (or organisations) have become members of the SNS [8], [10]. The information exchange may be done through instant messaging, emailing, voice recording, posting.

In this paper, we will attempt to design the “SignaComm”, a first SNS with an internationally oriented communication system for the protection of personal data on the Web. The SignaComm will be informative [8], [11]. It will execute two functions dynamically and in real time. Firstly, it will translate the member's input text into “signagrams” and deliver the result to another member. Secondly, it will display the history of instant messages in the chat room page. Our SNS would be used to deliver information to be understood and used quickly by its members. We hope that users from any culture, social environment or with disabilities could use it. The protection of personal data is defined by laws and regulations prohibiting the processing, storage or sharing of certain types of information about individuals without their knowledge and consent (e.g., analysing user's behaviour on a Website) [12].

Artificial communication systems: A number of artificial communication systems have been developed to improve the management of information, regardless of a specific natural language (e.g., Universal Playground, Istotype) [13], [14], [15]. We have been interested in the signage system, an artificial visual communication system with an international vocation where, the “signagram” is the writing unit [16]. The signagram's type is figurative as it is created from a direct representation of the object that evokes the object or situation to be represented [17]. Each signagram is made of an “external shape” (including the contours) and an “internal shape” [16] (Fig. 1).

The signage system and the signagram (the signage's unit) [16] will be integrated in our SNS, the SignaComm, to enable internationally oriented communication.

This paper's goal is to present the preliminary results of work in progress on the creation of the “SignaComm”. This SNS would support multilingual communication between users worldwide for the protection of personal data on the Web.

We design the SignaComm while relying on a theory and two principles: the theory of patterns [18], [19], [20], [21], as well as the principles of ontologies [22], [23], [24] and signage system [16]. At the core of Alexander's theory, a pattern describes the characteristics of a generic solution to a specific problem (e.g., the communication in real time between users worldwide). The theory of patterns permits the reuse and remodelling of patterns to serve as resources for software development and problem solving. According to Alexander [18, p. 313], “each pattern sits at the centre of a network of connections which connect it to certain other patterns that help to complete it”. The network of these relationships between small and large patterns creates the pattern. The ontology describes a structured set of concepts and objects by giving a meaning to an information system in a specific area (e.g., the user profile), and allows the construction of relationships between these concepts and objects [22], [23], [24].

The SignaComm could be implemented in the structure of a company's or public organisation's information system. Many fields may be interested in this SNS, for example, the cybersecurity, serious games, online learning. In our case, we are interested in the field of administrative authorities, namely the National Commission for Informatics and Liberty (in French, *Commission nationale de l'informatique et des libertés (CNIL)*) [25]. The CNIL is responsible for monitoring the data protection of professionals and individuals. We will explain the approach followed to develop the SignaComm for the protection of personal data when there may be a breach of privacy rights (e.g., the email advertising).

Our work consists of six sections. In section 2, we will present previously published works. In section 3, we will explain the SNS' characteristics and then design its pattern. In section 4, we will design the pattern for the automatic translation of text phrases into signagrams for the protection of personal data. In section 5, we will develop and test the prototype application that executes the SignaComm and communicates in visual messages, using the signage system and translation software of key-phrases into signagrams. In section 6, we will discuss the overall approach and finally conclude our work.

2 Related works

To our best knowledge, research projects addressing both topics, the SNS for multilingual communication and the protection of personal data, are limited. However, research projects are conducted on SNS combined with instant messaging and translation and, automatic translation of text phrases into signagrams. We will use our studies over these related works to fine-tune this research project and create the SignaComm.

In their published works, Yang and Lin [26] and Seme [27] have respectively developed a system and patent to automatically translate and send instant messages between members who communicate in different languages. Members, engaged in a session of instant messages, could send a message in a source language that could be translated automatically and received in a target language. The translation process has followed the Natural Language Processing (NLP) approach.

We have published works regarding a social networking site for crisis communication [28]. The objectives have been to translate in real time a sequence of syntagms into a series of signagrams, and to facilitate communication between members around the world. This SNS has translated automatically a source text into a target text (e.g., a message from the French language to the signage system) and has displayed the results in the SNS. The SNS has been based on the principles of the signage system, modular architecture and ontologies.

We have designed a software to automatically translate an input text into a sequence of signagrams [29]. We have relied on the semantic transfer method [30] with the linguistic rules and dictionaries for the source language and target communication system. The input has been the source text and has been written in the user's preferred language. The output has been the target text and has been written in the visual communication system, signage.

We rely on our works [16], [31] to show an example of signagram representing the syntagm *identify partners and data recipient* [25] (Fig. 1).

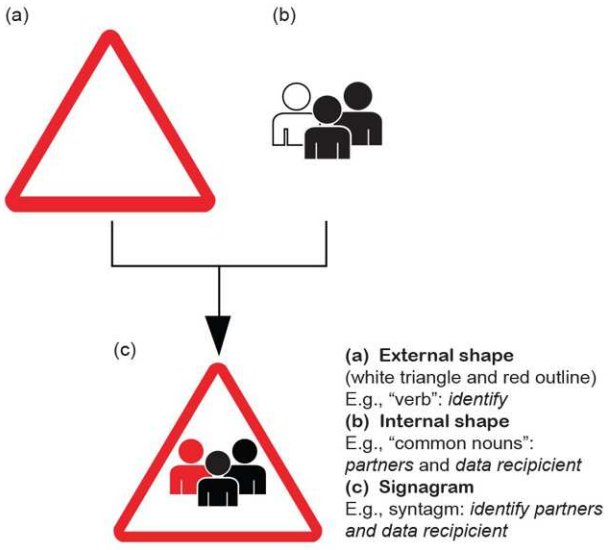


Fig. 1. Example of signagram.

3 Pattern for the SignaComm, first approach

We rely on the writings of Alexander [21] to create a pattern for the SignaComm [18]. This pattern is a first approach. The section consists of two main paragraphs, the descriptions of the SNS' context and design.

3.1 Description of the SignaComm's context

In general terms, the SNS interface follows the universal design principles of simplicity, flexibility and accessibility of use [32]. In addition, an SNS interface is graphical and contextual [33], [34]. Its graphical nature is based on a template that meets already defined and precise rules to ensure homogeneous and uniform results. These rules are the following: a simple and figurative content, uniqueness of graphical representations and uniqueness of colour contents [16], [28]. As for the dictionary of emoticons, it contains emotion symbols that are used worldwide.

So far, we haven't found published works regarding the standardisation of visual communication systems for SNS. A number of companies have developed their own communication system to integrate it into the SNS interface (e.g., Graphical User Interface (GUI) of WhatsApp) and the new technology tools (e.g., GUI of Apple iPhone). The company's (or organisation's) aim may be to intuitively guide users in their actions in entirely different and various contexts [33]. Each company (or organisation) chooses to adapt the charter of its visual communication system and its corresponding tools (e.g., SNS and new applications) according to the targeted countries. This adaptation approach may include the countries' laws, cultures, customs and traditions. Social media applications (GUI and emoticon dictionaries included) are essentially altered for two reasons. Firstly, to be compatible with international standards and regulations defined by every country government. And secondly, to meet the universal design principles depending on users' cultures.

To fulfil its objective, the SignaComm for the protection of personal data should utilise the signage system [16] as well as the graphical and contextual interface [33], [34]. The latter

should be meeting the universal design principles [32]. We have chosen both criteria to ensure the SignaComm's perception and spontaneous understanding worldwide. Furthermore, the SignaComm's development is in relation to three main concepts: the signage system [16], the SNS new technology tools [7] and user's adaptation. This interrelation makes the SignaComm dependent on these social, technical and human environments. For now, we will not include the emotional aspect as this SNS is an informative one.

3.2 Description of the SignaComm's pattern

The pattern for the SignaComm holds a network of connections between large and small patterns. In this work, we will present twelve large and two small patterns (in total 14 patterns) [19], [21]. The description is divided into several stages. A diagram will follow the explanations (Fig. 2).

We start with pattern 1 (Larger environments) that refers to many environments influencing the growth of SNS, such as information and communication technologies as well as social environments [6], [7].

Pattern 2 (Virtual communities environment) is contained inside pattern 1. Pattern 2 holds pattern 10 (SignaComm community) [6], [7].

Pattern 10 contains and describes the SignaComm functionalities (pattern 11), information technology administration (IT administration) (pattern 40) and interface (pattern 100) [3], [8].

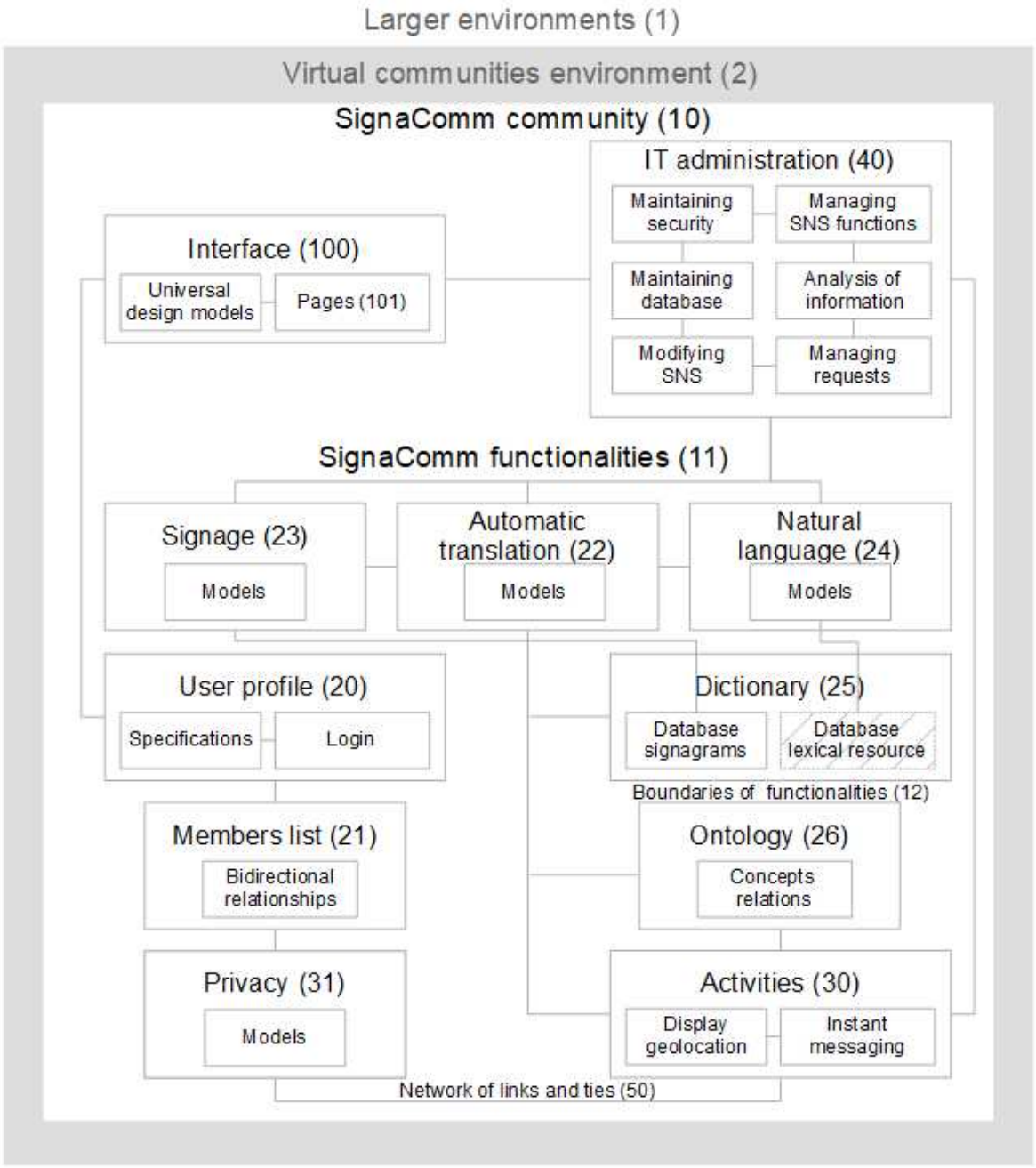


Fig. 2. Diagram for the SignaComm pattern.

Pattern 11 has various functionalities, listed as follow: the automatic translation of syntagms into signagrams (pattern 22), signage and signagrams models (pattern 23), natural languages and linguistic rules (pattern 24), dictionary (pattern 25), ontology (pattern 26), user profile characteristics and members' list (patterns 20 and 21), activities (pattern 30), privacy (pattern 31) [3], [8], [16], [30]. Every functionality has its own programming functions.

The SignaComm community (pattern 10) requires that all functionalities (pattern 11) execute their tasks to ensure the smooth running of the SNS. Pattern 12 (Boundaries of SignaComm's funtionalities) establishes boundaries to each functionality allowing it to perform its assigned tasks. It avoids the overlap with other functionalities.

Pattern 20 relates to user's profile characteristics [35], [36]. The SignaComm community encourages the diversity of members in order to enrich its growth [3]. Therefore, the growth of the SNS depends on a well balanced and represented community of members. This community would be able to support the interactions (pattern 30) between its members. For example, the interactions would help a member to solve a situation [3].

Pattern 20 has links with pattern 21 (Members list). The latter pattern specifies the SignaComm target audience (e.g., the bidirectional relationships) [37]. Members would belong to different cultures and social classes as well as different age groups [3].

Pattern 22 (Automatic translation) is a central functionality as it facilitates the communication between SignaComm members (pattern 21). Pattern 22 relies on the signage and signagrams, natural languages and linguistic rules, ontology as well as dictionary to translate members' requests (pattern 30).

Pattern 30 (Activities) is mostly linked to patterns 20, 21, 22 and 31 to create nodes of activities thus allowing members or groups to engage in various ways [7]. These activities may include invitations to join the SignaComm, instant messaging. Here, members have the opportunity to make acquaintances and connections, as well as to chat with members and groups of their choice [7]. Depending on the proximity of members, some ties are strong while others are weak [8], [10].

Pattern 31 (Privacy) is mainly for patterns 2, 20, 21, 22, 30 and 40. This pattern allows every member to set her/ his data sharing options with the IT administration, members and SNS environment [9], [12], [35], [36], [37], [38]. We provide the following example: a member chooses not to publicly display her/ his profile and then, not to share her/ his geographical position with the SignaComm and its environment. The SignaComm community (pattern 10) must respect the member's choice [12].

Pattern 40 (IT administration) is connected to both patterns 11 (SignaComm functionalities) and 100 (Interface). To make the interface and functionalities real, it is necessary to set up an IT administration. The latter manages the database and security, modifies the SNS, analyses the generated information and answers to members' requests.

Pattern 50 (Network of links and ties) creates and manages the network of relationships between all the patterns [7], [8]. It allows the information to circulate instantly and correctly in the SignaComm community.

Pattern 100 (Interface) gives an overview of the SignaComm interface, with an emphasis on the space of exchange between SignaComm members. The SNS' functionalities and IT administration contribute to its design [33], [34]. It includes the universal design principles [33].

Pattern 101 (Pages) is the continuation of pattern 100. The SignaComm is created with a reduced number of pages, such as the registration, members and chat room pages. This design is followed to quickly access information, provide flexibility in use and initiate intuitive interactions [32].

Fig. 2 shows an overall view of the SignaComm's pattern. It includes the fourteen patterns. We show the main links between the patterns to simplify the diagram's representation.

4 From text phrases to signagrams for the protection of personal data, first approach

Once we have designed the SignaComm pattern, we start developing pattern 22 (Automatic translation). As a reminder, the latter is a central functionality to achieve the SignaComm's objective. We rely on Emele et al. works and ours [16], [28], [29], [30] to accomplish this task. We will explain the methodology of work for developing both, the software and dictionary for the protection of personal data.

4.1 Automatic translation

We analyse the situation where a SignaComm member uses the application to translate in real time a sequence of syntagms (or text phrases) into a series of signagrams, and to engage in an informative conversation with a member or group of members. We present information to be quickly understood by members, to prevent some manipulation of personal data without their knowledge or permission and regardless of the computing device used [12], [25] (e.g., the portability of data).

While developing this machine translation prototype, we face a main difficulty, namely the non-figurative legal corpus. The suggested solutions are, on the one hand, to segment and analyse a thematic text and, on the other hand, to only translate the syntagms related to the case [29].

Here, for this machine translation, the expressions' exactness is necessary to be able to break down their relations with other encompassing units [29]. This would help by decreasing blunders and uncleanness in the translation process [39], [40]. Consequently, we utilise the National Commission for Informatics and Liberty portal's thematic text that presents reliable and relevant information [25].

Our model is composed of the ontology for the protection of personal data [41], [42], the construction of a dictionary of signagrams also related to the protection of personal data [43], [44] and the adaptation of the function translating text phrases into signagrams [28], [29], [30], [31].

We are particularly interested in the works of Palmirani et al. [41], [42] as their ontology is based on the application of the General Data Protection Regulation. The accuracy, flexibility and reliability of this ontology are well in line with our work objective. Therefore, it is appropriate to integrate it in the project.

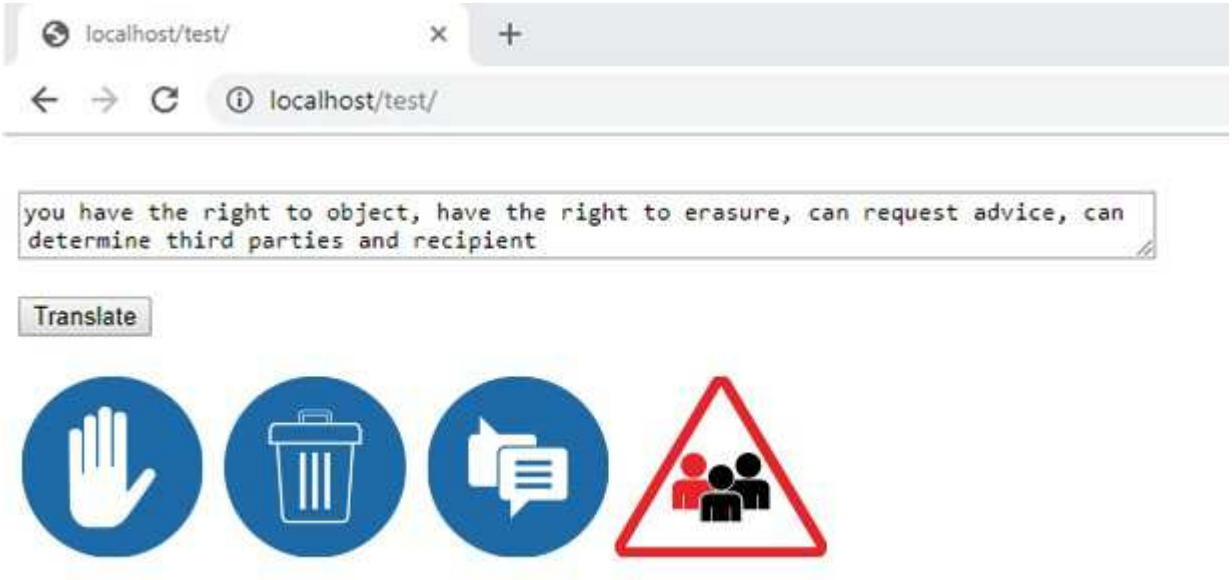


Fig. 3. Example of a machine translation result.

4.2 Dictionary of signagrams for the protection of personal data

To our knowledge, published works related to the dictionary of signagrams for the protection of personal data are limited. We rely on the works of Kordahi [45] and Takasaki [46] to design and develop this first dictionary, which is specialised. It provides information on signagrams to improve their understanding by any user.

The dictionary's design is based on the correspondence of vector signagrams to homologous semantic-based concepts [45], [46]. We program a mapping between two resources. The first semiotic graphical resource contains signagrams' external shapes (including the contours) and internal shapes [16]. The external and internal shapes, coming from that graphical resource, are stored in the dictionary. The second resource is a semantic lexical one (e.g., the WordNet [47]). The latter contains the concepts with their definitions and synonyms in English. The words, definitions and synonyms, coming from that lexical resource, are contained inside the dictionary.

We create fifty signagrams based on the works of Holtz et al. and the United Nations Economic Commission for Europe [43], [44], as well as the "Fotolia" international image bank [48]. The latter holds a large collection of images and symbols used globally. The signagrams' colours and shapes follow the international charter roads signs [44]. Fig 3 shows an example of automatic translation of text phrases into signagrams.

5. SignaComm's first technical test in the specific situation

For now, we have designed and programmed a prototype of the SNS. It is implemented in the Elgg platform and hosted on local and private servers.

The SignaComm is written with the PHP and Javascript programming languages to enable queries to be performed from a Web page. The interface is made up of a set of HTML Web pages. In this section, we choose to explain the four main patterns that are dynamically connected (section 3) [18], [19], [20], [21]. These patterns are the following: the interface (patterns 100 and 101), user profile (patterns 20 and 21), automatic translation of syntagms into signagrams (patterns 22 to 25) [28], [29], [45] and activities (pattern 30) [27].

5.1 SignaComm's interface pattern

The SignaComm's interface is used to display two sorts of information: the resulting information from an exchange between SNS members as well as interactions between the SNS system and its members. We provide the following examples, which include sending and receiving instant messages, displaying automatic translation of written texts into a sequence of signagrams, viewing a member's profile (section 3, Fig. 2).

The graphical user-interface consists of a main interface and secondary one. The main interface is used to display the Web pages' content. The secondary interface is the navigation bar. It enables the browsing between the various pages (Fig. 4).

5.2 SignaComm's user profile pattern

The user profile pattern performs three essential tasks. These are the registration of a user, invitation of a user and geolocation of members (section 3, Fig. 2). The first task allows a user to register and login to the SignaComm, which are a condition to use this SNS. The registration is done by submitting a user-name and password as well as some of the information regarding the user (e.g., choosing to share her/ his information [36], [37] and geographical position with the SNS) (Fig. 4). The login is done by submitting the member's user-name and previously saved password. The second function allows a SignaComm member to invite another user by sending an electronic invitation (e.g., instant message) while using the other patterns (e.g., pattern 21). This pattern 20 is connected to a geolocation process to allow performing the third task. The latter task automatically suggests a language of conversation [27].

This pattern comprises an application page and a PHP function. The HTML application page collects the user's registration information, including the name, physical address, email and address. The collected information is sent to the PHP function.

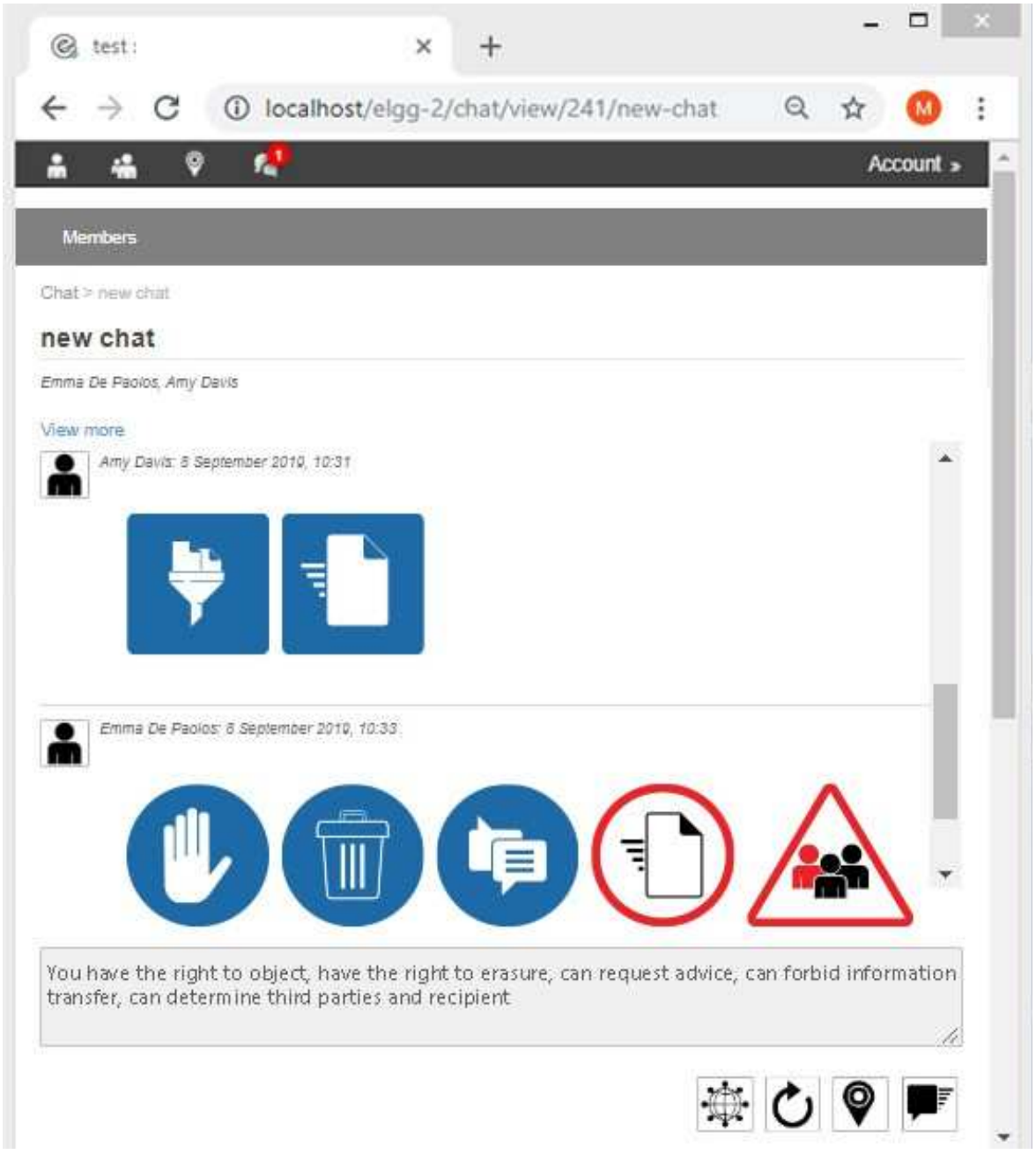


Fig. 4. Example of the chat room page.

5.3 Pattern of automatic translation

We rely on our works developed in sections 3 and 4 to implement the machine translation in the SignaComm structure.

Once implemented in the SignaComm, the translation pattern runs three consecutive tasks that are stored in this SNS database. The chat room page (written in HTML format) can receive the member's input text. A first request transmits the input text to be automatically translated into vector signagrams. A second request displays the machine translation result in the same HTML page. And a third request waits for the member's action to send the translated message to the activities pattern, or to reset the automatic translation process [28] (Fig. 4).

5.4. Pattern of activities

The pattern of activities performs two simultaneous and programmed tasks that are saved in the SignaComm. Chat histories are saved in the database's tables (section 3, Fig. 2). Through the server, the translation pattern receives requests from a member in the form of packets compliant with a common Internet protocol (e.g., the HyperText Transfer Protocol (HTTP) POST packets). These packets contain the translated information (the message is translated before delivery) [27]. The second task displays instant messaging exchange between members in the chat room page [27] (Fig. 4).

Fig. 4 shows an example of the SignaComm and the translation results. In this paper, the reported digital identities are simulated using fake profiles. Member 1 writes an input text (*we wish to collect information and transfer files*), activates its

automatic translation and then sends the resulting translation to a corresponding member 2. Member 2 replies to member 1 by writing, translating and sending a message [25]. The signagrams' reading direction is from left to right and top to bottom [15]. The result of Fig. 4 is comparable to Fig. 3.

6 Discussion and conclusion

While creating the SignaComm, with an internationally oriented communication system for the protection of personal data, we overcame at least one difficulty. To protect personal information on the Web, information accuracy, reliability, flexibility and speed of transmission are needed to assist individuals. We have formed the SignaComm of interrelated patterns. This interrelation has allowed us to synchronise the information exchange.

The obtained results demonstrate that the SignaComm is functioning correctly. In real time and instantly a sequence of text phrases is translated into a series of signagrams in order to send the results to members. Members can create their own network of contacts by inviting users of their choice. The geolocation process also identifies the member's preferred language.

Moreover, since the SignaComm for the protection of personal data is a first and a new prototype, we recommend preparing users for its utilisation with the aim to optimise its performance. This preparation should include detailed explanations regarding the SNS: the purpose, usefulness of its utilisation, interface functionalities and signage system. This preparation could be done in various ways, such as through a demonstration video, detailed guide, Questions and Answer (Q&A) forum. An online help would be interesting to design and implement in the SignaComm context. This would explain the SNS' social utility and the meaning of every signagram. Its use may be punctual, used to understand the meaning of a specific signagram or to search for a specific functionality.

In the near future, we would like to analyse and test the SignaComm with other writing systems, for instance Chinese. Furthermore, we wish to improve this first prototype. We will place the SignaComm in other areas and contexts, like the online learning one. In this context, on the one hand, we will analyse the digital identity of different SignaComm users/ members as well as the visibility models [36], [49]. On the other hand, we will conduct qualitative and quantitative studies on the user's behaviour while utilising the SNS. This study will allow us to evaluate, measure and improve the time required to understand a visual message.

Acknowledgment

We wish to thank Mohammad Haj Hussein, computer and communication engineer, for his valuable help while programming this prototype.

References

[1] A. Löwstedt, "Communication Ethics and Globalization," in *Ethics in Communication*, edited by Patrick Lee Plaisance, Volume 26 in de Gruyter Mouton Handbooks of Communication Science series, Peter Schulz and Paul Cobley, pp. 367-390, 2018.

[2] C. Alloing and J. Pierre, *Le Web Affectif: une Économie Numérique des Émotions,* INA éditions, 2017.

[3] B. Wellman, "Little Boxes, Glocalization, and Networked Individualism," In *Kyoto workshop on digital cities*, pp. 10-25. Springer, Berlin, Heidelberg, 2001.

[4] B. Wellman and M. Gulia, "Net-Surfers Don't Ride Alone: Virtual Communities as Communities," In *Networks in the global village*, Routledge, pp. 331-366. 2018.

[5] J. Fernback, "The Individual within the Collective: Virtual Ideology and the Realization of Collective Principles," *Virtual Culture: Identity and Communication in Cybersociety*, pp. 36-54, 1997.

[6] H. Rheingold, *The Virtual Community: Finding Commection in a Computerized World*, Addison-Wesley Longman Publishing Co., Inc., 1993.

[7] L. Raine and B. Wellman, *Networked: The New Social Operating System*. Massachussets: Massachusetts Institute of Technology, 2012.

[8] D. Boyd and N. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of computer-mediated Communication*, vol. 13, no. 1, pp. 210-230, 2007.

[9] H. Waheed, M. Anjum, M. Rehman and A. Khawaja, "Investigation of User Behavior on Social Networking Sites," PloS one, 12 (2), 2017, DOI:10.1371/journal.pone.0169693.

[10] M. Granovetter, "The Strength of Weak Ties: A Network Theory Revisited," *Sociological Theory*, vol. 1, pp. 201-233, 1983.

[11] C.M. Ma, Y. Zhuang and S. Fong, "Information Sharing over Collaborative Social Networks Using Xacml," In *8th International Conference on e-Business Engineering*, pp. 161-167, IEEE, 2011.

[12] E. Kennedy and C. Millard, "Data Security and Multi-factor Authentication: Analysis of Requirements under EU Law and in Selected EU Member States," *Computer Law & Security Review*, vol. 32, no. 1, pp. 91-110, 2016.

[13] T. Takasaki and Y. Mori, "Design and Development of a Pictogram Communication System for Children around the World," In *International Workshop on Intercultural Collaboration*, Berlin: Springer-Verlag, pp. 193-206, 2007.

[14] S. Fitrianie and L. Rothkrantz, "A Visual Communication Language for Crisis Management," I*nternational Journal of Intelligent Control and Systems* (Special Issue of Distributed Intelligent Systems), vol. 12, no. 2, pp. 208-216, 2007.

[15] M. Neurath, "Isotype," *Instructional science*, vol. 3, no. 2, pp. 127-150, 1974.

[16] M. Kordahi, "Signage as a New Artificial Communication System," *Canadian Journal of Information and Library Science*, vol. 37, no. 4, pp. 237-252, 2013.

[17] J.M. Klinkenberg, P*récis de Sémiotique Générale,* Louvain: De Boeck & Larcier, 1996.

[18] R. E. Kraut and P. Resnick, *Building Successful Online Communities: Evidence-based social design*, MIT Press, 2012.

[19] E. Gamma, R. Helm, R. Johnson and J. Vlissides, *Design Patterns : Element of Reusable ObjectOriented Software*, Addison-Wesley professional computing series, 1995.

[20] C. Alexander. *The Timeless Way of Building,* Vol. 1, New York: Oxford University Press, 1979.

[21] C. Alexander, *A Pattern Language: Towns, Buildings, Construction,* Oxford university press, 1977.

[22] T. Gruber, "A Translation Approach to Portable Ontology Specifications," *Knowledge acquisition*, vol. 5, no 2, pp.199-220, 1993

[23] N.F. Noy and D.L. McGuinness, "Ontology Development 101: A Guide to Creating your First Ontology", 2001.

[24] T. Gruber, "Collective Knowledge Systems: Where the Social Web Meets the Semantic Web," *Web semantics: science, services and agents on the World Wide Web*, vol. 6, no. 1, pp. 4-13, 2008.

[25] "National Commission on Informatics and Liberty",

hhttps://www.cnil.fr/en/home (15/01/20).

[26] C.Y. Yang and H.Y. Li, "An Instant Messaging with Automatic Language Translation," In *3rd IEEE International Conference on Ubi-Media Computing*, pp. 312-316, IEEE, 2010.

[27] Y. Seme, "Method and System for Translating Instant Messages," *U.S. Patent Application*, 10/035,085, filed July 3, 2003.

[28] M. Kordahi, "Réseau Social Numérique et Images Vectorielles: Introduction à une Communication à Vocation Internationale," *Recherches en Communication*, vol. 42, pp. 233-251, 2016.

[29] M. Kordahi and C. Baltz, "Automatic Translation of Syntagms into "Signagrams" for risk prevention", *Management des Technologies Organisationnelles*, vol. 5, pp. 23-37, 2015.

[30] M.C. Emel, M. Dorna, A. Lüdeling, H. Zinsmeister and C. Rohrer, "Semantic-based transfer," In *Verbmobil: Foundations of Speech-to-Speech Translation*, Berlin: Springer-Verlag, pp. 359-376, 2000.

[31] M. Kordahi, "La signalétique comme système de communication internationale, la protection des informations personnelles sur le Web," *International Association for Media and Communication Research (IAMCR)*, Madrid 2019.

[32] D. Spiliotopoulos, E. Tzoannos, C. Cabulea and D. Frey, "Digital Archives: Semantic Search and Retrieval," In *International Workshop on Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data,* pp. 173-182, Springer, Berlin, Heidelberg, 2013.

[33] R. Jain, J. Bose and T. Arif, "Contextual Adaptive User Interface for Android Devices," In *India Conference (INDICON)*, IEEE, pp. 1-5, 2013.

[34] D.B. Morin et al. "Method And System For A Personal Network," *U.S. Patent Application* 12/945,743, filed May 17, 2012.

[35] S. Proulx, "L'Irruption des Médias Sociaux: Enjeux Éthiques et Politiques," *Médias sociaux: enjeux pour la communication*, Québec : Presses de l'Université du Québec, 2012.

[36] D. Cardon, "The Design of Visibility, " *Réseaux*, vol. 6, no. 152, pp. 93-137.

[37] S. Bouraga, I. Jureta and S. Faulkner, "Requirements Engineering Patterns for the Modeling of Online Social Networks Features," In *4th international workshop on requirements patterns*, pp. 33-38. IEEE, 2014.

[38] R. Gross and A. Acquisti, "Information Revelation and Privacy in Online Social Networks," In *Proceedings of the ACM workshop on Privacy in the electronic society*, pp. 71-80, ACM, 2005.

[39] Y. Bar-Hillel, "The Present Status of Automatic Translation of Languages," *Advances in computers*, vol. 1, no. 1, pp. 91-163, 1960.

[40] M. McShane, S. Nirenburg and S. Beale, "An NLP Lexicon as a Largely Language-independent Resource," *Machine Translation,* vol. 19, no. 2, pp. 139-173, 2005.

[41] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini and L. Robaldo, "PrOnto: Privacy Ontology for Legal Reasoning"' *In International Conference on Electronic Government and the Information Systems Perspective*, Springer, Cham, pp. 139-152, 2018.

[42] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini and L. Robaldo, "Legal Ontology for Modelling GDPR Concepts and Norms," In *JURIX*, pp. 91-100, 2018.

[43] L.E. Holtz, K. Nocun and M. Hansen, "Towards Displaying Privacy Information With icons," In *IFIP PrimeLife International Summer School on Privacy and Identity Management for Life*, Springer, Berlin, Heidelberg, pp. 338-348, 2010.

[44] Economic Commission for Europe – Transport Division, Road Traffic and Road Signs and Signals Agreements and Conventions, https://www.unece.org/fileadmin/DAM/trans/conventn/Conv_road_signs_2006v_EN.pdf (15/01/20).

[45] M. Kordahi, "SignaNet: First specialised electronic dictionary for signage," *Pratiques et usages numériques H2PTM'2013,* pp. 385-387, 2013.

[46] T. Takasaki, "PictNet: Semantic Infrastructure for Pictogram Communication," in *The Third International WordNet Conference (GWC-06)*, pp. 279-284, 2006.

[47] G. Miller, *WordNet: An Electronic Lexical Database*, Cambridge, Massachusetts: MIT Press, 1998.

[48] "Fotolia", https://www.fotolia.com/ (15/01/20).

[49] S. Turkle, *Life on the Screen*, Simon and Schuster, 2011

Marilou Kordahi is a senior lecturer in information systems at Saint-Joseph University (Lebanon) and member of Paragraph laboratory, University of Paris 8 (France). She holds a PhD in Information and Communication Sciences for the University of Paris 8.

How useful are social networks for analyzing epidemics? The example of Tweeter for monitoring the 2018-2019 Ebola epidemic in Africa?

M. Tanti, K. Alate

***Abstract*— The last major Ebola outbreak in Africa occurred in 2014-2015. This epidemic has mainly affected three West African countries (Guinea, Sierra Leone, Liberia). She killed 20, 000 people. A number of articles have studied the rumors that circulated during this outbreak on Twitter. For example, Fung's article pointed out that these media disseminated false information about the treatment of the disease, such as bathing in salt water to heal (Fung, 2016). Jin's article also pointed out that these media were behind a fake news of a snake at the origin of the epidemic. This article also listed the Top10 rumors circulating on tweeter (Jin, 2014).**

Ebola virus disease has been raging in the Democratic Republic of Congo (DRC) since 1 August 2018. It killed more than 2050 people. It is the second largest epidemic after West Africa.

No studies have been conducted to determine who is communicating about this epidemic and what types of tweets/rumours are being disseminated?

To answer this question, we conducted an analysis on Tweeter via Radarly® software, over a period from April 1 to July 7, 2019. The keyword Ebola and #Ebola were used, with a filter on the French language. 17282 tweets were collected and classified via the software. The tool also extracted and represented the knowledge in a map-like way (volume of publications / time ...). It also identified the dominant themes in the form of clusters. The tone of the messages has also been determined.

Our work has highlighted several actors communicating around the epidemic: the general public, experts, politicians, the media ...

Concerning the general public, it was pointed out that the Congolese population communicating on the social network was shared. Some actors accepted the disease and acted as relays of preventive messages and fight against false rumors. But another party considered the disease to be a kind of conspiracy to destabilize the country and denied it. This part of the population spread false rumours and accused health workers of spraying the virus.

Concerning the experts, our study revealed that they shared many tweets to inform the general public.

Concerning the media/press, they are both Congolese and international. In particular, it was pointed out that they were disseminating prevention messages, including response teams.

Concerning the politicians, they supported medical teams by relaying educational messages.

Our study found 12 main influencers and revealed a negative message tone of 73.31%.

However, our study suffers from bias. We did not take into account other media such as Facebook and we limited our study to the French language.

***Index Terms*— tweeter; rumor; validated information; Ebola;**

I. INTRODUCTION

The last major Ebola epidemic in Africa took place in 2014-2015. This epidemic mainly affected three West African countries (Guinea, Sierra Leone, Liberia). It killed 20,000 people. A number of articles have investigated the rumors that circulated on Twitter during this epidemic. For example, Fung's article pointed out that these media disseminated false information about the treatment of the disease, such as bathing in salt water to cure. [1] Jin's article also pointed out that these media were behind fake news of a snake at the origin of the epidemic. This article also listed the Top 10 rumors circulating on tweeter (Figure 1) [2].

Fig. 1. Top 10 rumors circulating on tweeter [2].

Table 1. Top 10 Ebola-related rumors by Tweet volume from 28 September to 18 October 2014.

Rumor no.	Content	Label
1	Ebola vaccine only works on white people	White
2	Ebola patients have risen from the dead	Zombie
3	Ebola could be airborne in some cases	Airborne
4	Health officials might inject Ebola patients with lethal substances	Inject
5	There will be no 2016 election and complete anarchy	Vote
6	The US government owns a patent on the Ebola virus	Patent
7	Terrorists will purposely contract Ebola and spread it around	Terrorist
8	The new iPhone 6 is infecting people with Ebola	iPhone
9	There is a suspected Ebola case in Kansas City	Kansas
10	Ebola has been detected in hair extensions	Hair

Manuscript received 19th January, 2020.

M. Tanti, K. Alate work in the Centre d'épidémiologie et de santé publique des armées. Unité mixte de recherche 1252 – SESSTIM - Camp militaire de Ste-Marthe, 408, rue Jean Queillau, 13014 Marseille, France (corresponding author to provide phone: 0491637621; fax: 0491637825; e-mail: mtanti@gmx.fr.

Ebola virus disease is still raging today in the Democratic Republic of Congo (DRC) since August 1, 2018. It has killed more than 2,050 people. It is the second largest epidemic after that of West Africa [3; 4].

In addition, numerous studies have shown that Tweeter is used by public health organizations, in particular to inform, educate or monitor the state of health of populations, particularly in the event of a disaster [5]. However, no studies have been conducted to determine who communicates about the current Ebola epidemic in the Democratic Republic of Congo (DRC) and what types of tweets / rumors are circulated?

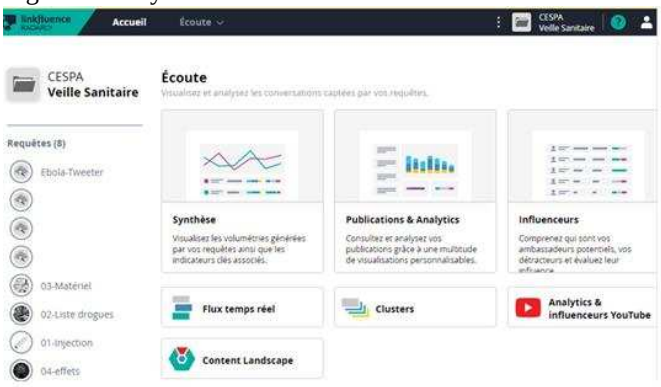
To answer this question, we conducted an analysis on Tweeter via the Radarly® software, over a period from 01/04/2019 to 07/07/2019. The keyword Ebola and #Ebola were used, with a filter on the French language. 17,282 tweets were collected and classified via the software. The tool also extracted and represented the knowledge in a cartographic way (volume of publications / time ...). It also made it possible to identify the dominant themes in the form of clusters. The tone of the messages has also been determined.

After a description of the methodology used, this article presents the main results found, in particular the fact that several actors communicate around the epidemic, in particular the general public, experts, politicians and the press.

II. METHODOLOGY

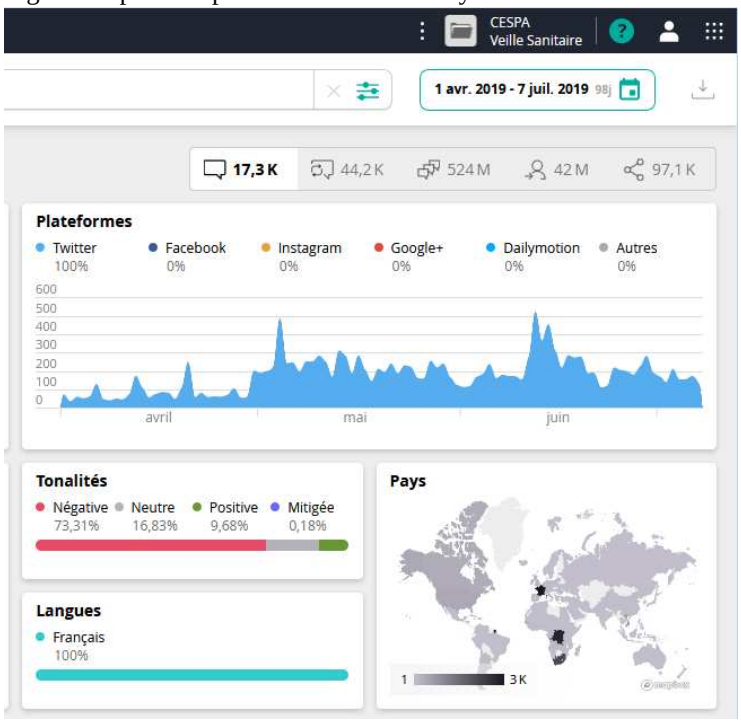
To carry out this study, we used the Radarly® social media monitoring software marketed by Linkfluence (https://radarly.linkfluence.com/login) and operating in SAAS mode (Figure 1). This software accessible online on subscription allows to collect data on the social web (Tweet, Facebook, Instagram, forums, blogs, etc.).

Fig. 2. Radarly® Interface



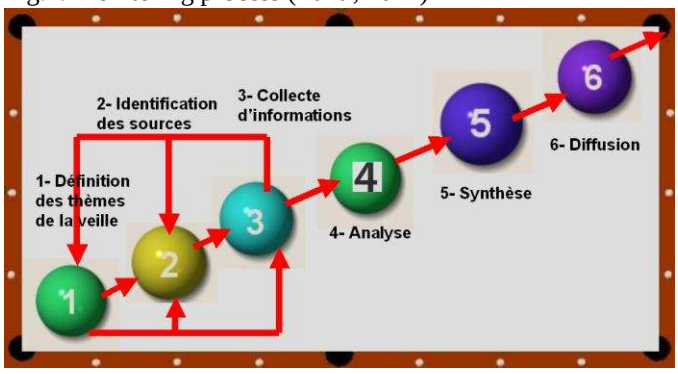
The software also makes it possible to represent the results in a cartographic manner, in particular in the form of clusters of dominant subjects. It also makes it possible to carry out analyzes of the tone of the messages published. It identifies "influencers" (people or groups who speak on a given topic or theme). It allows the export of data in .csv format to deduce statistics (Figure 3).

Fig. 3. Graphical representations of Radarly® data



To collect data from Radarly® software, we applied the monitoring process and the methodology (Figure 4) developed by Tanti in his article entitled "Pandémie grippale 2009 dans les armées : l'expérience du veilleur" [6] and which includes 6 stages: definition of monitoring themes, identification, collection, analysis, synthesis and distribution of documents.

Fig.4. Monitoring process (Tanti, 2012)



Concerning the first step, definition of the themes of monitoring, the keywords Ebola and #Ebola were used in the request, with a filter on the French language.

Concerning the second and third stages, identification and selection of documentary sources, we have only selected the collection of tweets on the social network tweeter via the platform.

Concerning the Analysis step, it was done using Radarly cartographic analysis and representation functionalities.

We have analyzed tweets posted on Tweeter only over a

period from 04/01/2019 to 07/07/2019. A total of 17,282 tweets were collected, classified and categorized via the software.

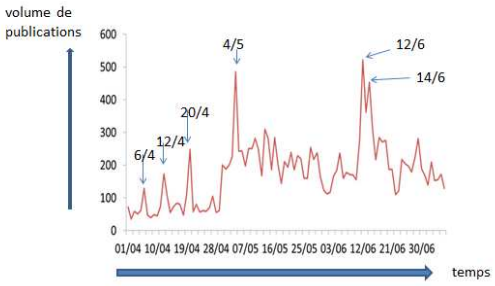
Figure 5 summarizes the number of data collected and analyzed (Figure 5).

Fig. 5. Data collected

- Nombre de publications : 17.282 publications.
- Nombre de publications et retweets : 44.166 publications et retweets.
- Nombre de reach estimé : 42 millions de reach estimé des publications. C'est le nombre estimé d'internautes ayant vu la/les publication(s) ou retweets
- Nombre d'Actions d'engagements : 97.122 actions : ceci est la somme des actions d'engagement sur chaque publication (likes, retweets, commentaires, partages, favoris…)

The software also enabled cartographic representations of the volume of publications as a function of time (Figure 6), making it possible to deduce media peaks intimately linked to health events.

Fig.6. Cartographic representation of the volume of publications function of time (01/04/2019 - 07/07/2019)



III. RESULTS

The requests have identified the actors who communicate on Tweeter concerning the Ebola disease which has been raging since 01/08/2018 in the Democratic Republic of Congo (DRC) and which has left more than 2,000 people dead.

The filter on the French language made it possible to select only tweets written in French. The analysis also made it possible to identify the messages conveyed.

The main players found are:

- The general public, mainly Congolese citizens and associations;
- Experts and health organizations (Ministry of Health of the Congo, WHO, etc.);
- The press, mainly Congolese…
- Politicians, mainly Congolese.

Concerning the general public

It is mainly Congolese citizens and associations who speak out on the epidemic on the social network. In the Congolese population, there are divided opinions and two populations: a population that « believes » and a population that does not « believe » in the disease.

The party that "believes" accepts the disease and the epidemic and considers it as a public health problem. It adheres to medical treatment and preventive measures. It relays messages of scientific information, education and awareness ... As an example, we can cite in Figure 7, a tweet relaying the effectiveness of the vaccine (Figure 7).

Fig 7. Tweet relaying the effectiveness of the vaccine

The rest of the population is less "gullible" and denies the disease. It constitutes the majority of the tweets found (high negative tone). Thus, despite the efforts of response teams since the start of the epidemic in August 2018, this population considers the disease as a plot to destabilize the country. This population spreads the rumor. She accuses, for example, the laboratories or the WHO of having created the virus (Figures 8).

Fig 8. Relay of a tweet accusing the laboratories

In the same context, Figure 9 shows a message spreading a rumor that the disease was sprayed from helicopters (Figure 9).

Fig.9. Tweet from a Congolese man spreading a rumor

Experts

It was mainly the national and international health organizations responsible for the response to the disease who

spoke on Tweeter during the study period. In particular, we observed that they shared many tweets to inform the general public. For example, the Ministry of Health of the DRC made a regular update on the disease (Figure 10) which it relayed on Tweeter.

Fig. 10. Tweet a daily newsletter from the Congolese Ministry of Health.

Nord-Kivu	Beni	415	9	424	261	270	62
	Biena	15	1	16	12	13	11
	Butembo	254	0	254	289	289	19
	Goma	0	0	0	0	0	24
	Kalunguta	125	15	140	53	68	15
	Katwa	607	16	623	420	436	27
	Kayna	8	0	8	5	5	5
	Kyondo	22	2	24	13	15	16
	Lubero	28	2	30	4	6	19
	Mabalako	338	16	354	237	253	14
	Mangurujipa	18	0	18	10	10	8
	Masereka	47	6	53	15	21	3
	Musienene	71	1	72	29	30	28
	Mutwanga	10	0	10	6	6	16

Media

It is mainly journalists and the press, both Congolese and international, who speak out. They tend to share WHO response releases, WHO prevention and awareness messages (Figure 11).

Fig. 11. Tweet from media relaying a preventive message1

Politicians

It is mainly Congolese politicians who speak out. They relay in particular prevention, health education or awareness messages. For example, Figure 6 shows a tweet relaying the photo of the President of the Congolese Republic (H.E.M. Felix Tshisekedi) who complies with medical requirements during his national tours (Figure 12).

Fig 12: Screenshot of a tweet showing the president of the DRC applying hygienic gestures

In conclusion, our study thus found 12 main influencers and highlighted a negative message tone of 73.31%.

IV. Conclusions

Our study, unlike the results found in the 2014-2015 epidemic, highlights a shared Congolese population. Some people accept the disease and adhere to the treatment and another party views it as a conspiracy.

Our study suffers from several biases. The short analysis period, like the choice to limit to the French language, is questionable. In addition, in our analysis, we did not take into account other media such as Facebook and forums. Finally, it is especially the choice of analysis on the social network Twitter itself which is questionable. Indeed, this media limits the number of characters present in messages. It thus limits long discussions, making it therefore the relay of current events and the engine of polemics and debates rather than the federator of true micro-communities.

In general, as a recommendation to change the mind of the defiant population, it would seem interesting to involve in the communication campaign on social networks, the relays of traditional healers and religious leaders who are the first to be consulted by this population.

.

References

[1] ICH. Fung & al, “Social Media’s Initial Reaction to Information and Misinformation on Ebola, August 2014: Facts and Rumors “, *Public Health Reports*, 131(3), pp. 461-473, 2016.

[2] F. Jin & al, “Misinformation Propagation in the Age of Twitter “, *Computer*, 47(12), pp. 90-94, 2014.

[3] Ebola Outbreak Epidemiology Team, “Outbreak of Ebola virus disease in the Democratic Republic of the Congo,

April-May, 2018: an epidemiological study. “, *Lancet*, 392(10143), pp. 213-221, 2018.

[4] AM. Medley & al, “Case Definitions Used During the First 6 Months of the 10th Ebola Virus Disease Outbreak in the Democratic Republic of the Congo - Four Neighboring Countries, August 2018-February 2019.“, *MMWR Morb Mortal Wkly Rep*, 69(1): pp. 14-19, 2020.

[5] M. Hart & al, “2017. Twitter and Public Health (Part 2): Qualitative Analysis of How Individual Health Professionals Outside Organizations Use Microblogging to Promote and Disseminate Health-Related Information“, *JMIR Public Health Surveill*, 3, e54. https://doi.org/10.2196/publichealth.6796, 2017.

[6] M. Tanti & al, “Pandémie grippale 2009 dans les armées : l’expérience du veilleur.“, *Médecine & Armées*, 40(5): pp. 389-401, 2012.

Deep learning of latent textual structures for the normalization of Arabic writing

Hammou Fadili

Abstract— Automatic processing of the Arabic language is complicated because of its many forms of writing, spelling, structure, etc., on the one hand, and the lack of preprocessed and normalized data, on the other. Implementing solutions that can help remedy these problems is a real need and a big challenge for the standardization process that this language must know, especially in the new world of publishing which is the Web; characterized by many forms of writing styles where everyone writes in his own way without any constraints. It is in this context that we propose an unsupervised approach based on deep learning implementing a system to help a normalization of a writing, according to a context characterized mainly by Arabic texts written in "Arabic" script.

Index Terms— Grammars and Other Rewriting Systems, Machine learning, Modeling and prediction, Natural Language Processing, Neural models, Normalization, Semantics

1 Introduction

The normalization of the Arabic language and of its writing are a necessity for its automatic processing. This mainly concerns the linguistic composition elements: syntactic, semantic, stylistic and orthographic structures. In order to contribute in this context, we propose an unsupervised approach based on deep learning implementing a normalization support system of writing in general, and in the context of written Arabic texts.

The approach is based mainly on the improvement of the Bi-RNN (Bidirectional Recurrent Neural Networks) model and its contextual implementation Bi-LSTM (Bidirectional Long Short-Term Memory) for sequences prediction. We implemented mechanisms to capture the writing signals encoded through the different layers of the network for optimizing the prediction of the next structures and the next words completing the sentences in writing progress or transforming texts already written into their normalized forms. The latent structures and the reference spelling are those learned from the training corpus considering their importance and their relevance: integration of the notions of attention and point of view in Bi-LSTM.

This article is structured as follows: in the first part, we present the learning model, in the second part, we define the methodological and technological elements implementing the approach. The third part is devoted to the generation of the dataset for learning. In the penultimate part, we present the tests and the results obtained, before concluding.

2 Motivation

The language writing normalization we are talking about concerns all aspects related to syntactic, semantic, stylistic, and orthographic structures. This is a complex task, especially in the case of languages used on the Internet, such as Arabic. Proposing solutions that can help authors and new users of such languages to respect and popularize their standards could be an important element in the process of standardizing their writings.

The rapid return and development of Artificial Intelligence makes such solutions possible: unsupervised approaches that do not require prerequisites or preprocessed data can be used to code the latent science contained in the studied corpora as a basis of machine learning and as references for normalization.

These are the elements that motivated us to study and propose an unsupervised approach allowing to detect and use the "NORMALIZATION" coded and contained in "well written" texts in terms of: spelling, composition and structures language.

Our contribution consists on the one hand in adapting and instantiating the contextual data model (cf. Hammou Fadili 2017) and on the other hand, in making improvements to the basic Bi-LSTM in order to circumvent their limits in the management of "contextual metadata".

• *Hammou Fadili*
Laboratoire CEDRIC du CNAM Paris
(Pôle Recherche & Prospective, Programme Maghreb) de la FMSH Paris
hammou.fadili(at)(cnam.fr , msh-paris.fr).

3 USED MACHINE LEARNING MODEL

Several studies have shown that deep learning has been successfully exploited in many fields, including that of automatic natural language processing (NLP). One of the best implementations is the generation of dense semantic vector spaces (Mikolov 2013).

Other networks such as RNN for Recurrent Neural Networks have also been improved and adapted to support the recurrent and sequential nature of natural language processing: each state is calculated from its previous state and new entry. These networks have the advantage of propagating information in both directions: towards the input layers and towards the output layers, thus reflecting an implementation of neural networks, close to the functioning of the human brain where information can be propagated in all directions by exploiting the memory principle (cf. the LSTM version of RNN in the following), via recurrent connections propagating the information of a ulterior learning (the memorized information).

These are the characteristics that allow them to take better care of several important aspects of natural language. Indeed, they have this ability to capture latent syntactic, semantic, stylistic and orthographic structures, from the order of words and their characteristics, unlike other technologies such as those based on the concept of bag of words (BOW) where none order is not considered, obviously involving loss of the associated information.

In serial RNNs, each new internal state and each new output simply depends on the new entry and the old state.

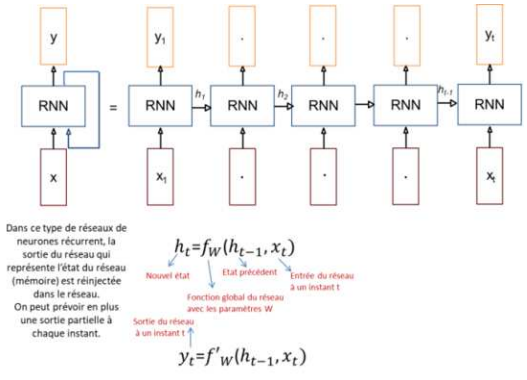


Fig 1. How an RNN works

RNNs can also be stacked and bidirectional, and the previous simple equations can be redefined for the two learning directions according to the model below.

$$Vers\ l'avant:\ h_t^l = \tanh W_t^l \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^l \end{pmatrix}, h_t^l \in \mathbb{R}^n, W^l matrice\ de\ dimension\ [n \times 2n]$$

$$Vers\ l'arriere:\ h_t'^l = \tanh W_t'^l \begin{pmatrix} h_t^{l-1} \\ h_{t+1}'^l \end{pmatrix}, h_t^l, h_t'^l \in \mathbb{R}^n, W'^l matrice\ de\ dimension\ [n \times 2n]$$

$$Sortie:\ y_t = \tanh W_t \begin{pmatrix} h_t^l \\ h_t'^l \end{pmatrix}, y_t \in \mathbb{R}^n, W^l matrice\ de\ dimension\ [n \times 2n]$$

The training of multi-layers RNN is done, as for the other types of networks, by the minimization of the error (difference between the desired output and the output obtained) which one obtains by the back-propagation of the error and the descent of the gradient. It can be demonstrated mathematically that the depth of RNNs which can be high, because of sequential nature of paragraphs, texts, documents, etc., generally depends on the number of words to be processed at a time; can provoke:

- Either the Vanishement of Gradient in the first layers and the end of learning from a certain depth. When we must multiply the gradient, a high number of times, by a weight w / $|w|<1$.
- Either the Explosion of Gradient always in the first layers and the end of learning from a certain depth. When we must to multiply the gradient, a high number of times, by a weight w / $|w|>1$.

The LSTM architecture makes it possible to remedy these problems (Hochreiter, S., & Schmidhuber, J. (1997). It is based on finer control of the information flow in the network, thanks to three gates: the forget gate which decides what to delete from the state (h_{t-1}, x_t), the input gare which chooses what to add to the state and the output gate which chooses what we must keep from the state (cf. equations below).

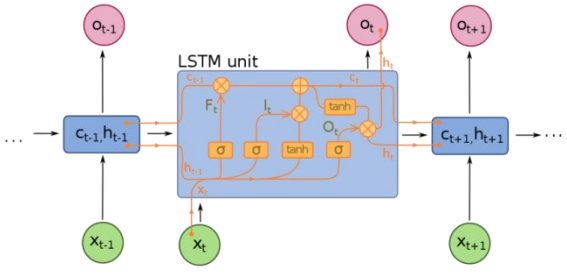


Fig 2. How a basic LSTM works

$$F_t = \sigma(W_F x_t + U_F h_{t-1} + b_F) \quad \text{(forget gate)}$$
$$I_t = \sigma(W_I x_t + U_I h_{t\ 1} + b_I) \quad \text{(input gate)}$$
$$O_t = \sigma(W_O x_t + U_O h_{t-1} + b_O) \quad \text{(output gate)}$$
$$c_t = F_t \circ c_{t-1} + I_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = O_t \circ \tanh(c_t)$$
$$o_t = f(W_o h_t + b_o)$$

These equations defining the learning process of an LSTM express the fact that this kind of network allows to cancel certain useless information and to reinforce others having a great impact on the results. We can also show by mathematical calculations that this architecture allows, in addition to the optimization of the calculations in the network, to solve the problems linked to the vanishing and the explosion of the gradient. This is what motivated our choice to use and improve this model by adapting it to our needs. We also integrated the notion of perspective or point of view of analysis as well as the notion of attention in the general process.

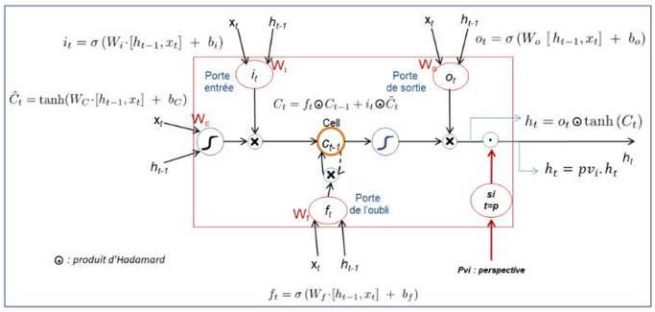


Fig 3. Architecture of an improved LSTM

So, our model allows to control the flow in the "Context" i.e.:

- What to forget from the state
- What to use from the state
- What to send to the next state
- According to a point of view or perspective
- By paying attention to relevant information

This latter notion is represented outside the internal architecture of the LSTM model.

4 TECHNOLOGICAL ELEMENTS INTERGRATION

Our approach has two main objectives:

- Help users / authors to normalize their writing during the writing process

- Normalize already written texts as an early preprocessing step in an automatic analysis

In the first case, our approach aims to predict and suggest supplementing the sentences being written with the most relevant words in their standardized orthographic forms and following a very specific linguistic structure (learned from texts of the training corpus). In other words, to a sequence of words that the user is typing, the system proposes him a sequence of words, sentences or parts of standardized sentences.

In the second case, our approach aims to normalize the already written texts as an early pretreatment step in a semantic automatic analysis process. This by transforming the texts, sentences and words into their normalized forms according to the orthographic, linguistic, stylistic structures, etc. learned from the training corpus.

In both cases, this requires the use of the sequence-to-sequence or seq2seq version of Bi-LSTMs.

The proposed architecture consists of the following main layers:

- Encoding
  - Pretreatment
  - Internal representation
  - Domain (Point of view)

- Decoding
  - New internal representation (calculated)
  - Attention
  - Prediction

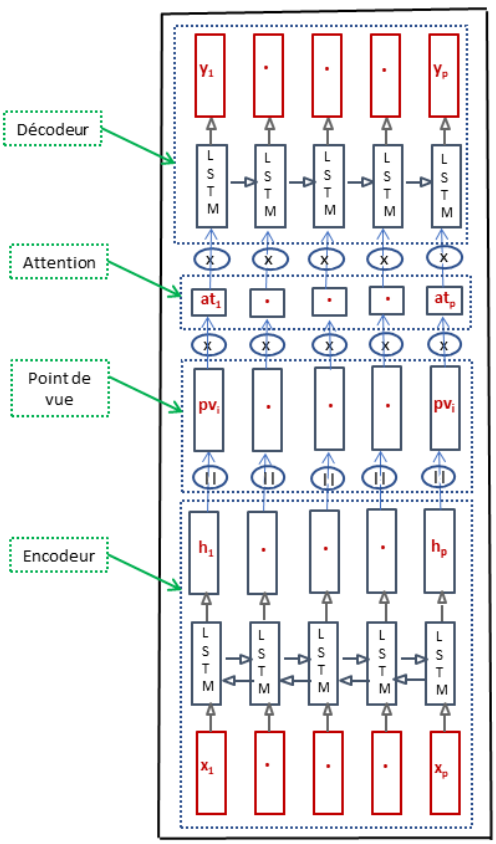


Fig 4. Architecture

Une perspective est un ensemble de mots caractérisant un point de vue d'analyse (espace multidimensionnel).
Soit p le nombre de mots de la phrase courante
Soit i l'indice de perspective courante
Soit P une perspective définie par les dimensions $p_1, p_2, \dots, p_l$
Soit H la matrice composée par les représentations cachées $h, h_2, \dots, h_p$ générées par le LSTM
Considérons $pv_i = \sum_1^p d_j$ *où* $d_j = \sum_1^l d_k$ *et* $d_k = \frac{p_k . h_j}{\|p_k\| \|h_j\|}$

pv_i est une première pondération dans le calcul de la prédiction. Elle est basée sur la somme des distances cosinus entre les représentations cachées générées par le LSTM et les dimensions de la perspective d'analyse ciblée.

Cela permet d'obtenir une première transformation, pondérée par le rapprochement à la vue considérée, des représentations internes des sens de la phrase :

$H' = (h'_1, h'_2, \dots, h'_p)$, *où* $h'_j = d_j . h_j$
d_j *est la somme des ditances cosinus de hi à chaque dimension.*
$\propto_d = softmax(\sum_1^p at_i)$
$at_i = w_i^T . h'_i$ *constitue l'attention apprise par* (w_i) *pour chaque* h'_i

$prediction = H \propto_d^T$ *représente la sortie du système.*

7 Corpus and dataset

The corpus was built up by the collection of many documents in Arabic, obtained mainly from "well-written" institutional websites. The learning data model was obtained from a simplified version of the language model and the extended semantic model (Hammou Fadili. 2017) and projections of the initial vector representations of the words, relating to a space of large dimensions (vocabulary size), in a reduced dimensions semantic space using Word2vec technology (w2v). The goal is to create an enriched model of instances adapted to the context of language normalization, with a reasonable size vector representation, essential for optimizing calculations and processing. The instantiation of the model was done by splitting the texts into sentences, and the generation of the enriched n-grams context windows for each word:

- 2-grams
- 3-grams
- 4-grams
- 5-grams.

This first version of the instances is enriched with other parameters to support the syntactic structures and the thematic distributions. We associated each word in the dataset with its grammatical category (Part Of Speech POS) as well as its thematic distribution (Topics) obtained by LDA (Latent Dirichlet Allocation) (Bei & all. (2003)).

The other latent structures have been integrated by the coding of the sequential nature of the texts by the LSTM. Similarly, the spelling has been coded from the training corpus. It is this dataset that is provided to our augmented and improved Bi-LSTM model.

This model has the advantage of considering, in fine, the local context (n-grams) and the global context (themes), long-term and word order memories which encode the latent structures of the texts (syntactic, semantic, etc.).

8 Experiences & evaluations

We have developed several modules and programs, implementing the elements of the general process, including mainly:

- Automatic extraction of the various characteristics (Features).
- Implementation of a deep learning system based on Bi- LSTM endowed with the attention and domain mechanisms.

We have also designed an environment centralizing access to all modules:

- Integration of all the elements in a single Workflow implementing all the processing modules.

Features extraction:

We have developed and implemented 3 modules that run in the same "python-Canvas" environment and exploited a process based on a "General MetaJoint" of the same environment to make the link between the different components.

Without going into details, the 3 modules consist, in the order, of extracting the following characteristics:

- The first allows to extract the "linguistic context" of each word through a sliding window of size *n*
- The second allows grammatically annotating all the words in the text: perform POS Tagging (Part Of Speech Tagging)
- The third implements the "Topic modeling and LDA" technology in order to automatically extract the studied domain(s).

Workflow:

The implementation of the Workflow was done in the Orange-Canvas platform. It automates the chaining of results from previous modules for the extraction of characteristics and the instances ready to be used for learning.

Improved Bi-LSTM neural network:

We developed from scratch a personalized Bi-LSTM neural network and adapted it to our needs. It is a neural network (multi-layer & recurrent Bi-LSTM) endowed with attention and domain mechanisms.

In order to ensure the honesty of the system, we have separated the generated learning data into three parts:

- A first part for validation on 20% of the dataset, in order to optimize the hyper-parameters of the system: the learning step, the type of the activation function and the number of layers.
- The rest of the dataset divided into two parts:
  - 60% for training, in order to estimate the best coefficients *(wi)* of the function of the neural network, minimizing the error between the actual outputs and the desired outputs.
  - 20% for tests, to assess the performance of the system.

During all the learning phases, the system is autonomous, it can generate the features of the text (Features) for training and perform the learning process (validation, training and tests).

Résultats

```
=== Evaluation on test split ===
Time taken to test model on test split: 0.75 seconds
=== Summary ===
Correctly Classified Instances       455          83.7937 %
Incorrectly Classified Instances      88          16.2063 %
Kappa statistic                  0.8323
Mean absolute error              0.0021
Root mean squared error          0.0302
Relative absolute error         21.9222 %
Root relative squared error     44.2078 %
Total Number of Instances        543
```

After training the system on the constituted corpus, we conducted tests on a small text of 543 words. The system succeeded in correctly predicting the location and the ponctuation of 455 words in the different sentences of the text, considering the latent structures coded from the training texts. The system, however, failed on the remaining 88 words and ponctuation.

For more information on reading these results cf. (read Weka) in the references.

9 Conclusion

The conducted experiments confirm the contribution, in good ratio, of the proposed approach, to correct and help the normalization of Arabic texts.

As described above, we have adjusted several hyper-parameters to optimize the system and generate all the parameters of the instances. On the entire corpus obtained, we validated the approach on 20% of the corpus, trained the system on approximately 60% and tested on approximately 20%. The various measurements show the good performance of the approach, in terms of accuracy and loss.

This is mainly due to improvements made to the integration of certain aspects of the language and to extensions implemented for Bi-LSTM:

- Language model and its instantiation
- Introduction of the concepts of attention, domain and point of view in Bi-LSTM.

References

[1] A Neelakantan, J Shankar, A Passos, A McCallum. Efficient non-parametric estimation of multiple embeddings per word in vector space. Conference on Empirical Methods in Natural Language Processing, 2014

[2] Cui Tao, Dezhao Song, Deepak Sharma, Christopher G. Chute, Semantator: Semantic annotator for converting biomedical text to linked data. Journal of Biomedical Informatics, Volume 46, Issue 5, Pages 882-893 (October 2013). DOI: 10.1016/j.jbi.2013.07.003

[3] Das, T. K., & Kumar, P. M. (2013). Big data analytics: A framework for unstructured data analysis. International Journal of Engineering and Technology, 5(1), 153-156.

[4] Boury-Brisset, A.-C. (2013), Managing Semantic Big Data for Intelligence., in Kathryn Blackmond Laskey; Ian Emmons & Paulo Cesar G. da Costa, ed., 'STIDS' , CEUR-WS.org, , pp. 41-47 .

[5] Delia Rusu , Blaž Fortuna , Dunja Mladenić. Automatically Annotating Text with Linked Open Data (2011).Venue: In 4th Linked Data on the Web Workshop (LDOW 2011), 20th World Wide Web Conference.

[6] Archit Gupta, Krishnamurthy Viswanathan, Anupam Joshi, Tim Finin, and Ponnurangam Kumaraguru. Integrating Linked Open Data with Unstructured Text for Intelligence Gathering Tasks. Proceedings of the Eighth International Workshop on Information Integration on the Web, March 28, 2011.

[7] Isabelle Augenstein. Lodifier: Generating Linked Data from Unstructured Tex". ESWC 2012

[8] Marin Dimitrov. From Big Data to Smart Data. Semantic Days May 2013

[9] Khalili, A.; Auer, S. & Ngonga Ngomo, A.-C. (2014), conTEXT – Lightweight Text Analytics using Linked Data, in 'Extended Semantic Web Conference (ESWC 2014)'.

[10] E. Khan, "Addressing Big Data Problems using Semantics and Natural Language Understanding," 12th Wseas International Conference on Telecommunications and Informatics (Tele-Info '13), Baltimore, September 17-19, 2013.

[11] E. Khan, "Processing Big Data with Natural Semantics and Natural Language Understanding using Brain-Like Approach", submitted to Journal– acceptance expected by Dec. 2013 Jan 2014.

[12] James R. Curran, Stephen Clark, and Johan Bos (2007): Linguistically Motivated Large-Scale NLP with C&C and Boxer. Proceedings of the ACL 2007 Demonstrations Session (ACL-07 demo), pp.33-36.

[13] Hans Kamp (1981). A Theory of Truth and Semantic Representation. In P. Portner & B. H. Partee (eds.), Formal Semantics - the Essential Readings. Blackwell. 189-222.

[14] Minelli, Michael & Chambers, Michele & Dhiraj, Ambiga 2013. Big Data, Big Analytics: Emerging Business Intelligence and Analytics Trends for Today's Businesses.

[15] Chan, Joseph O. "An Architecture for Big Data Analytics." Communications of the IIMA 13.2 (2013): 1-13. ProQuest Central. Web. 6 May 2014.

[16] H. Fadili. Towards a new approach of an automatic and contextual detection of meaning in text, Based on lexico-semantic relations and the concept of the context., IEEE-AICCSA , May 2013.

[17] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[18] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[19] Hiemstra, P.H., Pebesma, E.J., Twenhofel, C.J.W. and G.B.M. Heuvelink, 2008. Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network. Computers & Geosciences, accepted for publication.

[20] Christian Bizer, Tom Heath, Kingsley Idehen, Tim B. Lee. Linked data on the web (LDOW2008), In Proceedings of the 17th international conference on World Wide Web (2008), pp. 1265-1266.

[21] Jianqing Fan, Fang Han, Han Liu. Challenges of Big Data analysis National Science Review, Vol. 1, No. 2. (1 June 2014), pp. 293-314.

[22] http://wiki.dbpedia.org/

[23] Publication MEDES 2016 : Towards an Automatic Analyze and Standardization of Unstructured Data in the context of Big and Linked Data. H. FADILI.

[24] Publication TICAM'2016 : Le Machine Learning : numérique non supervisé et symbolique peu supervisé, une chance pour l'analyse sémantique automatique des langues peu dotées. H. FADILI.

[25] Frijda, N. H., Mesquita, B., Sonnemans, J., & Van Goozen, S. (1991). The duration of affective phenomena or emotions, sentiments and passions.

[26] Shand, A. F. (1920). The foundations of character: Being a study of the tendencies of the emotions and sentiments. Macmillan and Company, limited.

[27] Zhou, X., Tao, X., Yong, J., & Yang, Z. (2013, June). Sentiment analysis on tweets for social events. In Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on (pp. 557-562). IEEE.

[28] Park, C., & Lee, T. M. (2009). Information direction, website reputation and eWOM effect: A moderating role of product type. Journal of Business research, 62(1), 61-67.

[29] Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2014). Text mining for market prediction: A systematic review. Expert Systems with Applications, 41(16), 7653-7670.

[30] Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013, January). Mining online user-generated content: using sentiment analysis technique to study hotel service quality. In System Sciences (HICSS), 2013 46th Hawaii International Conference on (pp. 3119-3128). IEEE.

[31] Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. Icwsm, 10(1), 178-185.

[32] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

[33] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. Computational linguistics, 37(2), 267-307.

[34] Denecke, K. (2008, April). Using sentiwordnet for multilingual sentiment analysis. In Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on (pp. 507-512). IEEE.

[35] Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML-14) (pp. 1188-1196).

[36] Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. IEEE Intelligent Systems, 28(2), 15-21.

[37] Agarwal, A., Biadsy, F., & Mckeown, K. R. (2009, March). Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 24-32). Association for Computational Linguistics.

[38] Besag, J. (1986). On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society. Series B (Methodological), 259-302.

[39] Paltoglou, G., & Thelwall, M. (2012). Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media. ACM Transactions on Intelligent Systems and Technology (TIST), 3(4), 66.

[40] Singh, V. K., Piryani, R., Uddin, A., & Waila, P. (2013, January). Sentiment analysis of textual reviews; Evaluating machine learning, unsupervised and SentiWordNet approaches. In Knowledge and Smart Technology (KST), 2013 5th International Conference on (pp. 122-127). IEEE.

[41] Rao, D., & Ravichandran, D. (2009, March). Semi-supervised polarity lexicon induction. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 675-682). Association for Computational Linguistics.

[42] Esuli, A., & Sebastiani, F. (2007). SentiWordNet: a high-coverage lexical resource for opinion mining. Evaluation, 1-26.

[43] Hung, C., & Lin, H. K. (2013). Using objective words in SentiWordNet to improve sentiment classification for word of mouth. IEEE Intelligent Systems, 1.

[44] Boudia, M. A., Hamou, R. M., & Amine, A. (2016). A New Approach Based on the Detection of Opinion by SentiWordNet for Automatic Text Summaries by Extraction. International Journal of Information Retrieval Research (IJIRR), 6(3), 19-36.

[45] Forrester. (2016). Think You Want To Be "Data-Driven"? Insight Is The New Data. [online] Available at: https://go.forrester.com/blogs/16-03-09-think_you_want_to_be_data_driven_insight_is_the_new_data/.

[46] Amiri, H., & Chua, T. S. (2012, July). Sentiment Classification Using the Meaning of Words. In Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.

[47] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.

[48] Zhang, W., & Skiena, S. (2009, September). Improving movie gross prediction through news analysis. In Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology-Volume 01 (pp. 301-304). IEEE Computer Society.

[49] Zhao, J., Dong, L., Wu, J., & Xu, K. (2012, August). Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1528-1531). ACM.

Hammou Fadili is a HDR (Habilité à Diriger des Recherches) researcher at Conservatoire National des Arts et Métiers (CNAM) in computer science. He is a member of the information systems team, named Ingénierie des Systèmes d'Information et de Décision (ISID). Also, he is a manager of the Digital Humanities projects at the research department of the FMSH and associate researcher at the Digital Humanities department of the Paris 8 university. His research work concerns especially these areas: Semantic Web, WEB 3.0, Machine learning (supervised, unsupervised, etc.), Digital Humanities, Automatic Natural Language Processing (NLP), Language & Context Modeling, Detection and anonymisation of sensitive data, Semantic mining of structured and unstructured data, (Linked, Big, Smart) data. He is a member of the program committee and member of the organization committee for various conferences. He is also reviewer for some journals. He is (or was) involved in many ANR (Agence Nationale de la Recherche) and European research projects.

An Ontology-based recommender system for Service-Oriented Dynamic Product Lines

Najla Maalaoui, Raoudha Beltaifa, and Lamia Labed Jilani

Abstract—Service-Oriented Dynamic Software Product Line (SO-DSPL) is a rapidly emerging software reuse development paradigm for the design and development of adaptive and dynamic software systems based on service-oriented architecture. In this framework, features are software systems functional requirements that allow stakeholders performing the configuration of multiple products. This is done through the interactive selection of a valid combination of features which are mapped to services. However, since stakeholders are often unsure about their needs, focusing on relevant configuration options becomes challenging. Likewise, in a SO-DSPL context, the configuration must be adaptable to anticipate any changes. Thus, the configuration activity must not only take into account the requirements of the clients, but also it must consider different factors such as customers' profiles variability, contexts variability and services dependencies variability in order to take into consideration the dynamicity of the systems.In fact, they must anticipate to the adaptations caused by different types of contexts. To help stakeholders and to improve the efficiency and quality of the product configuration process, we propose to recommend service-products to users based on a recommender system framework. However, recommender system in the context of a SO-DSPL framework is different from the typical recommender systems due to the variability, contexts changes and the constraints which must be taken into account during the recommendation activity. This paper describes a SO-DSPL recommender system meta-model based on OntoUML conceptual modeling. This is done in order to present different dimensions of the recommender system framework in order to manage and capitalize the necessary knowledge to build the most suitable recommendations.

Index Terms—Recommender system, Service-Oriented Dynamic Software Product Line, ontology.

1 Introduction

The evolution of service computing has allowed organizations to focus on more complete software applications. However, these applications became more complex in order to satisfy maximum customer's needs. Offering to the customers a panel of options among which they can select their preferred ones is a way to defining customer's needs. To satisfy different customer's needs, it is necessary to provide the ability of producing customized products (such as software or systems) based on methods, techniques and tools engineering. In order to achieve this, software product lines are used in order to develop a family of products sharing common characteristics and differing in some variability points to satisfy the needs of a particular mission [1]. Within the SPLE field, Dynamic Software Product Lines (DSPLs) have emerged as a promising mean to develop reusable and dynamically reconfigurable core assets.According to this evolution, Service-oriented dynamic software product lines (SO-DSPL) are defined as a class of DSPLs that are built on services and service-oriented architectures (SOAs) [2]. From a product line engineering viewpoint, the SO-DSPL deployment model promises two major benefits over the traditional software deployment model with a fixed configuration: 1) the dynamic nature of SOAs means that SO-DSPL can support user needs and expectations in a continuously changing environment and 2) a SO-DSPL can combine services in various configurations and contexts, simplifying the deployment of product variants that are mostly based on the same core services but tailored to meet different customers' needs. To describe commonalities and variabilities of the software product line, feature model has been used. Feature models allow users to select a valid combination of features which cover user requirements. This activity is known by "product configuration "activity, which refers to the decision-making process of selecting an optimal set of features from the product line that comply with the feature model constraints and fulfill the product's requirements [4] . In practice, choosing from a wide range of options becomes quickly difficult for the customer who doesn't know where to start, or which alternative to choose. Besides each times a customer makes a decision, this can be contradictory with previous decisions, or have a negative impact on downstream decisions. Therefore, it is crucial to guide the customer in the SPL configuration process.

The configuration activity is more crucial in the context of SO-DSPL because of variability management, context awareness and service dependencies caused by their reuse in various situations and contexts. Indeed, customers are often unsure about their needs in one hand, and they do not have knowledge about services mapped to features, possible context changes and their impacts on the customized configuration on the other hand. Since these features are represented by web services, the choice of a feature remains a crucial task. In fact, the customers can have knowledge about the descriptive data of the service but context-sensitive information remains restricted to the service provider. Particularly, in SO-DSPLs a bad selection of services is not only relative to the customer, but it is

• *N. Maalaoui,R. Beltaifa and L. Labed were with with Laboratoire de Recherche en Génie Logiciel, Applications distribuées, Systèmes décisionnels et Imagerie intelligentes (RIADI), Université de Manouba et Université de Tunis.*
E-mail: najla.maalaoui@ensi-uma.tn, raoudha.beltaifa@ensi.rnu.tn, lamia.labed@isg.rnu.tn

also relative to the product line environment, where the process of dynamic adaptation or runtime reconfiguration [3] can encounter contradiction problems, which produce problems of dissatisfaction with constraints. To deal with this problem, we propose to recommend a configuration based on different data, such as previous configuration, user preferences, user profile, services descriptions and other data.However,the choice of data to be used as a dataset is considered as the key of the success of a recommender system. Thus, we propose in this paper an ontology that capture useful data to a SO-DSPL recommender system from different views. The conceptual modeling of our proposed ontology is based on OntoUML language. OntoUML is an ontologically well-founded language for Ontology-driven Conceptual Modeling. It is built as a UML extension based on the Unified Foundational Ontology (UFO) [21]. The proposed SO-DSPL recommendation ontology emphasizes the semantics carried by the different elements of ontology, as well as the semantic relationships between them,in order to perform recommender system knowledge management.
The remainder of this paper is organized as follows. Section 2 presents a background about SO-DSPL and the recommender systems. Section 3 provides an overview of the related works, which discusses service selection approaches in SO-DSPL.In section 4,we present our proposed approach and in Section 5 we evaluate it.In section 5,we summarize our contributions and discuss the perspectives for further work.

2 Background

In this section, we introduce product-line engineering with a brief overview of the domain and application engineering phases. Furthermore, we present the basic concepts about recommender systems.

2.1 Software product line

Software Product-line engineering is a paradigm within software engineering, used to define and derive sets of similar products from reusable assets [1]. The development life-cycle of a product line encompasses two main phases: domain engineering and application engineering [1] . While domain engineering focuses on establishing a reuse platform, application engineering is concerned with the effective reuse of assets across multiple products.
Domain engineering:
Domain engineering is responsible for defining all common and variable assets of a product line and their respective interdependencies. In practice, feature models are the main formalism to represent and manage reusable assets, called features, and their interdependencies [1] . A common visual representation for a feature model is a feature diagram.A feature diagram is a tree-based structure, where each node represents a feature, and different edges illustrate the dependencies between two or more features [1] .The feature diagram defines common features found in all products of the product line, known as mandatory features,and variable features that introduce variability within theproduct line, referred to as optional and alternative features.
Application engineering:
Application engineering is responsible for capturing the requirements of a product, defining a selection of features that fulfill these requirement and comply with the feature model's dependencies, and deriving the product from this selection of features [1] . A suitable selection of features (i.e. configuration) results from the product configuration process. Due to the combinatorial possibilities of selecting variable features, the number of possible configurations can grow exponentially with the number of features in a product line. Therefore, configuration processes are often incremental such that a user starts with an empty configuration and selects or deselects one feature at a time until the configuration is valid and fulfills the product's requirements. Nevertheless, considering the large configuration space for even product lines with a relatively low number of features, finding the best configuration for given product requirements can still be a challenging task.

2.2 Service oriented- dynamic software product line

The limitations of SPL rely on its inability to change the structural variability at runtime, provide the dynamic selection of variants, or handle the activation and deactivation of system features dynamically and/or autonomously [3].As variability becomes more dynamic, SPL approaches moved to recent development approaches like Dynamic Software Product Lines as an emerging paradigm to handle variability at runtime and at any time. Dynamic software product lines are able to dynamically switch an executing system from one variant to another, without stopping its execution, without violating its feature model's constraints and without degrading its functional and non-functional behavior. Consequently, the feature model itself must become a runtime entity that can dynamically evolve to satisfy new variability dimensions in the system.
Since Service Oriented Architecture (SOA) has proven to be a cost-effective solution to the development of flexible and dynamic software systems [3], the combination of SOA and DSPL can provide significant mutual advantages. DSPL can provide the modeling infrastructure required to understand a running SOA-based system.In particular,these models can be used to understand the implications of modifying a system's configuration at runtime. The combination of SOA and DSPL has led to the service oriented dynamic software product lines.

2.3 Recommender system

The rapidly evolution of the market environment has led to the development of recommendation systems for large companies like Amazon, YouTube, Netflix and eBay. A recommendation system is a personalized information filtering technique used to suggest a set of elements that will most likely interest a particular user [5]. The suggestions provided are intended to guide the user through various decision-making processes, such as the product line configuration process. This recommendation process is based on information filtering algorithms to predict whether a particular user has the intensity of liking a particular feature or product. Recommender systems use learning algorithms that can find relevant items from a large set of items in a personalized way [5]. The most common classes of such

algorithms are collaborative filtering algorithms, content-based recommender systems and Knowledge-Based recommender systems.

Collaborative Filtering (CF) algorithms are based on relevance feedback from users. The feedback has the form of ratings (e.g., a 5-star rating for relevant items and a 1-star rating for non-relevant items) that are stored in a user item-rating matrix. CF algorithms can be divided into two categories. The first of them is the neighbourhood-based CF . Neighbourhood-based CF algorithms search for neighbours of the active user in the user-item-rating matrix. After determining the neighbourhood of the active user, predictions of ratings for items unknown to the active user are made by calculating weighted mean rating of all neighbours. We adapt this method to recommending features in productline configuration.

The second class of CF algorithms is Matrix Factorization (MF). MF algorithms decompose the user-item-rating matrix approximately into two latent matrices. The decomposition is done by minimizing an error function using e.g., stochastic gradient descent. The latent matrices are used then to make rating predictions of unknown items. MF algorithms have shown their great predictive power in numerous studies [5].

Content-based recommender systems analyze the content of items. Then, given the history of a user, they search for items similar to the ones the user purchased before. Contentbased systems are largely used in scenarios in which a significant amount of attribute information is available at hand. In many cases, these attributes are keywords, which are extracted from the product descriptions. knowledgebased recommender system makes recommendations based on specific queries made by the user. It might prompt the user to give a series of rules or guidelines on what the results should look like, or an example of an item. The system then searches through its database of items and returns similar results [5].

knowledge-based recommender system makes recommendations based on specific queries made by the user. It might prompt the user to give a series of rules or guidelines on what the results should look like, or an example of an item. The system then searches through its database of items and returns similar results [5]. Knowledge-based methods require the explicit specification of user requirements to make recommendations, and they do not require any historical ratings at all. Therefore, these methods use different sources of data, and they have different strengths and weaknesses.

3 RELATED WORK

In recent years, Service selection and recommendation have been extensively studied to facilitate Web service composition. Earlier studies were interested in functionality-based Web service recommendation that refers to recommending services by matching user requests with service descriptions. However, using keyword-based matching usually suffers from poor retrieval performance. Therefore, many semantics based approaches had been proposed such as [6] and [7]. These approaches leveraged domain ontologies and dictionaries to enrich descriptions of both services and user requests, and adopted logic-based reasoning for semantic similarity calculation. Nevertheless, this kind of approaches is difficult to apply to large-scale service data because ontologies are manually defined. For this reason, on one hand, some researches explore machine learning and data mining technologies in functionality-based service recommendation. For example, In [8], the authors proposed an approach for mining domain knowledge on service goals from textual service descriptions based on linguistic analysis. On the other hand, many web service recommendation approaches based on Collaborative Filtering (CF) are proposed. CF refers to recommending services according to the past composition history, the similarity of users, or the similarity of services. Some approaches used CF to predict quality of service (QoS), which can be used to select high quality services. For example, Jieming et al [9] proposed an online QoS prediction for runtime service adaptation approach based on an adaptive matrix factorization. Recently, many approaches combine machine learning and collaborative filtering to recommend services such as [14], [11] and [12]. Ruibin et al. [14] proposed a deep hybrid collaborative filtering approach for Web service recommendation.

In [11] a QoS web services prediction approach based on k-means clustering was proposed. Fuzan et al. [12] proposed a web service discovery approach using semantic similarity and clustering. In SO-DSPL, proposals are limited to services selection at runtime, such as in [13] and [10]. Jackson et al. [13]proposed an approach for services selection in order to bind DSPL features that should be connected to a product at runtime. The services are selected based on an analysis of the services quality attributes. Benali et al. [10]proposed a software framework which supports context-awareness behavior to assign services to consumers based on DSPL features .

Based on the studied works in the literature, we conclude that the recommendation of the services in SO-DSPL remains a challenge until now.To meet these challenges, we propose an an ontology for recommender system in SO-DSPL framework in order to conceptualize relevant data for the recommendation process.

4 OFSOPLR: ONTOLOGY FOR SO-DSPL RECOMMENDER SYSTEM

OfSoPLR is an ontology for a recommender system in SO-DSPL framework.The proposed ontology describes common conceptualization of knowledge to make suitable recommendation in the context of SO-DSPL. Our proposed ontology is built with the main goal of making the best possible description of the needed knowledge for generating recommendations.In particular,OfSoPLR is designed to be used as a core conceptual model to be used for several purposes, such as:

- As a reference model for annotating recommender systems data in a semantic documentation approach. Once semantically annotated, we can extract knowledge and link contents from different data which can be useful for recommendation process.
- To support human-learning about key concepts related to SO-DSPL configuration via recommendation-based configuration process.

Our research goal is to explore the combination of two complementary forms of guidance: recommendation and configuration. In order to cover this scope, OfSoPLR should be able to answer the following competency questions:

- CQ1:What type of data should be used for recommendation at domain engineer? (experiences of past configurations, profiles, services, etc)?
- CQ2: how the data are semantically linked to each other?
- CQ3: What type of data should be used for recommendation at application engineering ? (partial configurations, profiles, contextual data, etc)?
- CQ4: How to combine configuration and recommendation?
- CQ5: how recommender systems can benefit from SO-DSPL architecture?
- CQ6: which modules must be available in a recommender system at domain engineering to perform a good recommendation?
- CQ7: which modules must be available in a recommender system at application engineering to perform a good recommendation?
- CQ8: Which recommendation technique shall be used? (collaborative filtering, content-based filtering, etc)

As mentioned earlier, combining configuration process with recommendation process provides significant mutual advantages. Many companies try to provide solutions for customization of their products to meet the needs of each user. However, these solutions need to consider a large set of variabilities.In software product line framework, configurator tools do not yet adequately support business needs [15], which is a more crucial problem in SO-DSPL framework due to the dynamic context changes and the large number of services that involved in the runtime adaptation process.As mentioned in [16], after an interview with eight IT experts, the following three main configuration challenges are faced by the company when using configurators are identified:
Challenge 1: large product lines. Industry product lines often define several thousand features. Thus, the amount and complexity of the options presented by the configurator lead IT experts to get lost with so much information to be taken into account, and spend too much time reasoning about too many options and complex relationships. Thus, showing all features and their dependencies is impractical as a decision-maker can only focus on one part of the configuration a time.
Challenge 2: unclear requirements and features. The requirements are very often not clear and exact due to the vague descriptions used to the requirement specification. Moreover, the feature model may present many subjective features that cannot be matched with the requirements.
Challenge 3: get a final configuration. It may be very difficult to define a valid configuration since quite often stakeholders specify requirements that are inconsistent with the feature model's constraints.
From these three challenges, we deduce the following challenges relative to the context of the SO-DSPL framework:
Challenge 4: get a contextual configuration. It is very difficult to define a configuration not based only on the requirement of the user, but also based on user, service line and environment contexts. In most cases, configuration can satisfy all user needs but because of a context change (e.g. a service down, problem in data center, user skills are not consistent with the knowledge necessary to use the services), the configuration will be not valid at the current time. Also, a contextual configuration must be valid in any instant t, however, Since service-oriented systems needs adaptation at runtime environment, a configuration must have an available set of adaptation rules that will be used in a context change scenario. In some cases, a configuration needs adaptation caused by a context change, however, adaptation rules can not be executed because of constraints dependencies (e.g necessity of using an international payment service in a country which can only provide national bank cards to its people).
Challenge 5: select the relevant data to be considered in the configuration process . It is difficult to decide which data must be considered to get the best contextual configuration.In SO-DSPL, considering only feature model in a configuration process is not enough.Indeed, services,contexts and the configuration algorithms must be considered.However,he way how to use this data is challenging and differs from a situation to another.
In summary, business experiences show that the product configuration process can be challenging as decision makers regularly do not know every feature, their interdependencies and their influencer contexts, particularly for large service-oriented product lines.

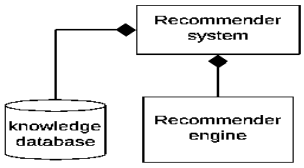


Fig. 1. Recommander system basic components

Every recommender system is composed of two basic components, namely: recommender engine and knowledge data base as shown by Fig.1.
Our proposed ontology is divided into two sub ontology, namely:OfSoPLR Domain and OfSoPLR application. Each sub-ontology is articulated around three-dimensions for recommender's data modeling, namely user, SO-DSPL and service dimension as presented by fig 2. The OfSoPLR ontology is used by the recommender system presented by the two meta-models Domain recommender dimension and Application recommender dimension.

4.1 OfSoPLR Domain

In the domain engineering level, the presented three sub-ontologies aim to model all possible information that can be collected and used by the recommender system . For example, the user dimension aims to represent all possible users information that can be useful for the recommendation process, independently from the recommendation method. At application engineering level, recommendation

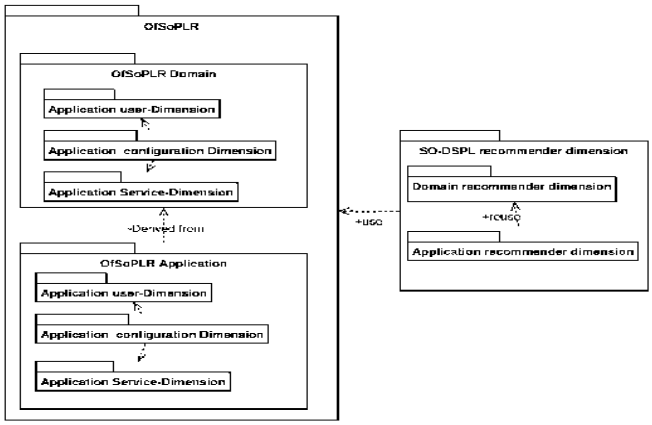


Fig. 2. OfSoPLR sub-ontologies

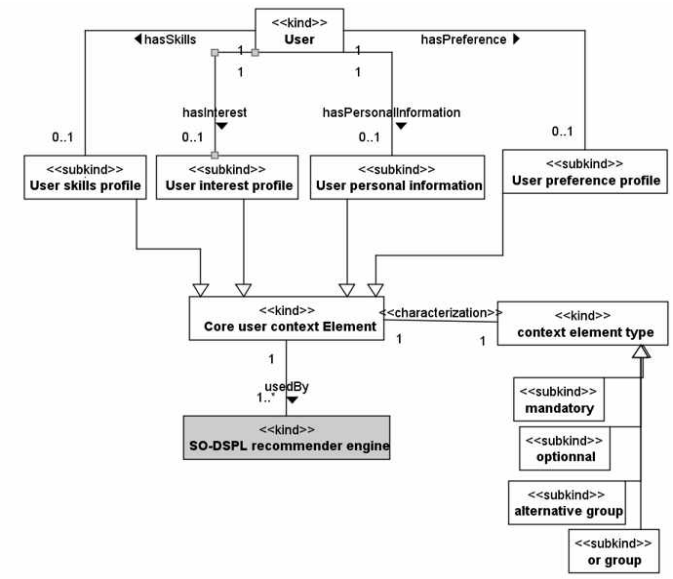


Fig. 3. OfSoPLR domain user dimension sub-ontology

methods will be chosen based on the current situation, user information to be considered will be selected, configuration and there correspond services will be derived based on the selected elements from the domain engineering. The three dimensions presented by each sub-ontology aims to cover respectively user view, software product line view, services view.In order to present the way of using these sub-ontologies by the recommender system in domain level,a domain recommender system meta-model is defined.
The basic recommender's information at domain engineering are conceptualized from three dimensions which are presented by the following sub-ontologies.

4.1.1 OfSoPLR domain user dimension sub-ontology :

In a context aware, the context information is used to respond and adapt to users' behaviors, which is the case of a contextual recommendation. Indeed, considering user profile and its context information at recommendation process make recommendation more contextual and more adaptable to the user context change and needs. In order to consider this aspect, a user dimension sub-ontology has been presented with the aim of consolidating and standardizing the ontological user resources derived from an exhaustive study and analysis of existing contributions in user profile modelling,which can make solution to the challenges 4 and 5.
This sub-ontology addresses the competency questions CQ1 and CQ2. In [18] is shown an ontology to represent and capture the user profiles within a changing environment, where the user model is considered as a type of the context. We use this user profile ontology to model user context at domain level, as shown in Fig.3.The concept core user context element presents a kind of user context, a core user context element can be divided into user profile, user interest profile, user personal information profile, user preference profile and user skills profile. To relate user with his/her context element, semantic relations are predefined. Thus, the user concept is related to the concept User skills profile by the relation "hasSkills", to the concept User interest profile by the relation "hasInterest", to the concept User personal information profile by the relation "hasPersonal-Information" and to the concept User preference profile by the relation "hasPreference".

4.1.2 OfSoPLR domain SO-DSPL dimension sub-ontology

In the context of a recommendation in the SO-DSPL, the interaction between the configurator and the product line is managed through its model, and more specifically its feature model. Thus, we devote in our ontology a dimension dedicated to the SO-DSPL. This sub-ontology addresses the competency questions CQ1 and CQ2. We use ontological feature model representation proposed in [20] enriched by knowledge that we have extracted from the literature, as shown in Fig.4.A SO-DSPL is composed of features that can have one or more attributes.A feature can be a "mandatory feature","optional feature"," alternative feature" or "or group feature".According to the dynamicity of the dynamic SPL, since each feature can be activated or deactivated at runtime, a feature has a possible activation mode that can be divided into active and disable mode. The management of the feature's dynamicity is based on constraints, so the possible activation mode has constraint to describe the conditions of the activation and the deactivation of the feature. A feature can have three possible variability types [3], namely: temporal variability, spatial variability and contextual variability, which are considered as sub-kind of variability type.
The spatial variability implies configurable functionality of a software system in terms of different features dening the set of all possible configurations. The contextual variability refers to capture influences of the environment on functionality of software systems. The temporal variability refers to capture evolution of software systems. As old versions of software systems need to be supported with updates, it is important to keep track of the evolution of SPLs. Spatial, contextual and temporal variability strongly correlate. Contextual variability changes the set of selectable conguration options of the spatial variability. Temporal variability captures evolution of spatial and contextual variability. Consequently, the type of variability and their evolution influence the validity of the recommended configuration.The SO-DSPL uses core user context element in order to perform

a contextual configuration at application engineering and it is used by the SO-DSPL recommender engine to generate initial recommendation.

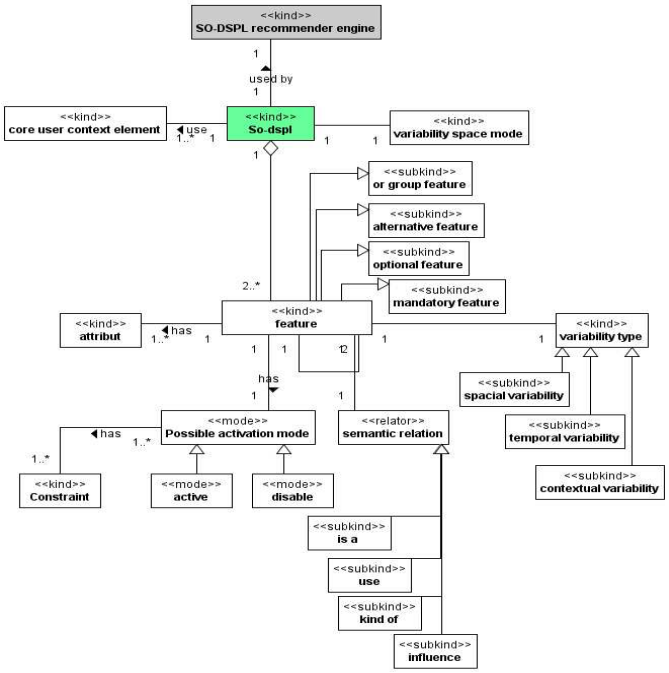


Fig. 4. OfSoPLR domain SO-DSPL dimension sub-ontology

4.1.3 OfSoPLR domain service dimension sub-ontology

The OfSoPLR domain service dimension sub-ontology addresses the competency questions CQ1 and CQ2.Our ontological service representation is based on the OWL-S ontology. OWL-S is an ontology language within the OWL-based framework of the Semantic Web, for describing Semantic web service. This ontological representation enables users and software agents to automatically discover, invoke, compose, and monitor Web resources offering services, under specified constraints [19].OWL-S ontology has three main components which are: the service profile, the process model and the grounding as shown in Fig. 5; the service profile is used to describe what the service does, the service model describes how a client can interact with the service and finally the service grounding specifies the details that a client needs to interact with the service, as communication protocols, message formats, etc.

Our service dimension representation is also enriched with other knowledge extraction from literature and organization feedbacks,as shown in Fig.6. The concept "Core service" represents service at domain engineering level, likely to owl-s ontology, the core service has a service profile,a service model and a service Grounding. A core service can be atomic or composite, where a composite service is composed of Core Service. A core service has many service instances which provides the same functionality with different ways and QOS, it has also a criticality level that reflects how important it is to the software product line operations, this level can be high, low or medium.The core service and all it related concepts are usedBy the SO-DSPL recommender engine.To define the constraint that

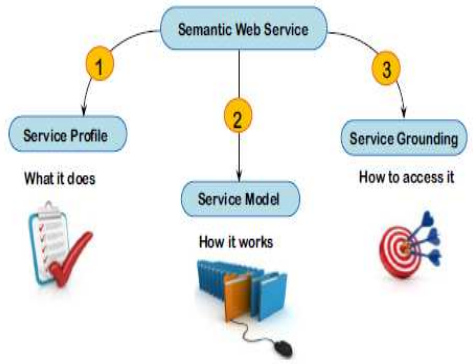


Fig. 5. Semantic web service

all concepts related to the core service are automaticely used by the SO-DSPL recommender engine, we created an abstract relation namely "ServiceRelation" where all relation presented in this sub-ontology are sub-relations and we define the axiom A1.To answer CQ2, a relationship between SO-DSPL's feature and Core Service is introduced, each feature of the SO-DSPL is mapped to at last one Core service that covers its functionalities.To represent this constraint, axiom (A2)is defined.

(A1) SC : CoreService; SRC : SODSPLRecommender; t :Thing usedBy(SC,SRC) $\wedge$ $ServiceRelation(SC,t) -> usedBy(t;SRC)$

$(A2) \forall SPL : SO-DSPL, F : Feature, composedOf(SPL,F) -> \exists CS : Core-Service, coveredBy(F,CS)$

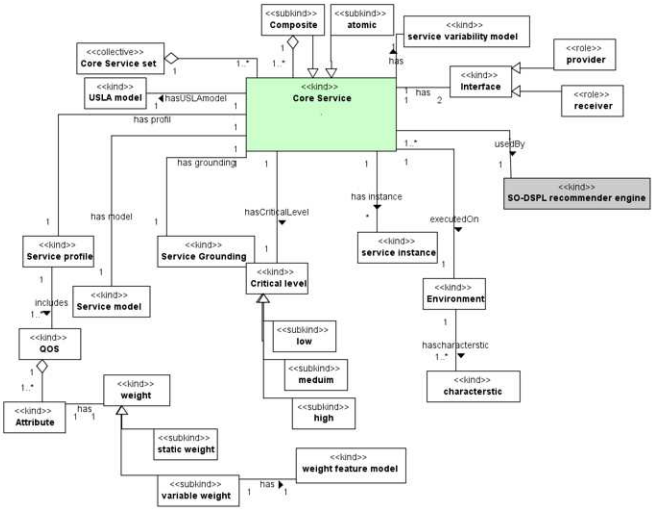


Fig. 6. OfSoPLR domain service dimension sub-ontology

A core service has a service variability model which indicates the variability of service instance, for example,a purchase service can be executed with different inputs input1, input2, input3,input4, if the input2 is chosen, input 4 must be specified and security quality of service will be considered, but if the input 1 is chosen, not other input must be specified and availability QOS will be considered. The service profile includes QOS which are composed of

attributes. These attributes have different weights that can be static or variable. A static weight is fixed by the service provider however a variable weight changes according to the context and the situation.Each core service has two interfaces, namely: provider interface and receiver interface. Each core service is linked to a feature of the feature model and can be executed in one or more environment that is characterized with one or many characteristics.

4.1.4 Domain recommender system meta-model

As mentioned above, at this level the recommender engine uses data from domain and application engineering to prepare initial recommendations, analyze previous situation to decide which recommendation method is more suitable and train used model such as neural network models. To ease this process, we propose the domain recommender engine dimension sub-ontology to conceptualize used data and needed modules to perform recommendation process at domain engineering. This meta-model addresses the competency questions CQ1,CQ2,CQ4,CQ5,CQ6 and CQ8.

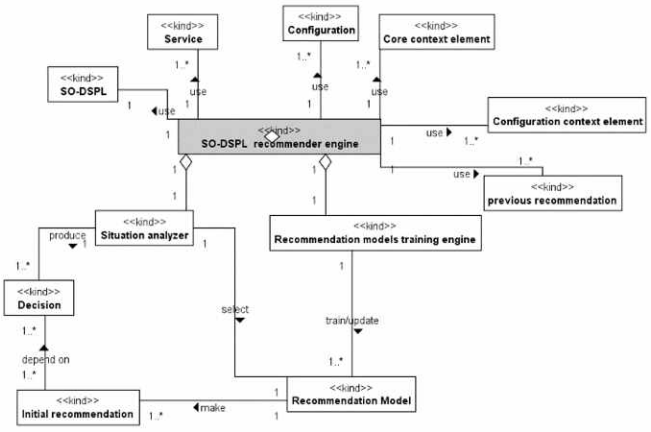


Fig. 7. Domain recommender dimension meta-model

To perform recommendation, as shown in Fig.7, recommender engine use data capitalized by service sub-ontology, SO-DSPL sub-ontology, previous recommendations that can be performed at application engineering or prepared in domain engineering , configuration derived from the SO-DSPL at application engineering, application user context dimension sub-ontology which present the monitored user context at an instant also, recommender engine at domain engineering is composed of modules, namely: Situation analyzer, model selector engine and Recommendation model training engine. The situation analyzer module aims to analyze the available data and contexts to recommend users and decide which recommendation method or model will be more suitable to prepare initial recommendation.This decision is based on predefined metrics. All situations and taken decisions are capitalized to be used in the next iteration. The Recommendation models training engine aim to train, update and test available recommendation models using application engineering recommendation results.

4.2 OfSoPLR Application

At domain engineering, all possible recommendations' data are modeled, many recommendation situations are analyzed, recommendation models are trained and partial configuration(recommendation) are prepared. At application engineering, it is the time to recommend a configuration to the current user following his request. Based on the available data that are selected from the domain engineering domain, recommender engine at this level uses results from the domain recommender engine and produces relative ones.Thus,the OfSoPLR Application sub-ontology is composed of three sub-ontologies that are derived from domain ones, namely: OfSoPLR application user dimension sub-ontology,OfSoPLR configuration dimension sub-ontology,OfSoPLR application service dimension sub-ontology and used by a recommender system presented by an application recommender dimension meta-model.The relation of derivation designates the selection of the concepts to be used from the domain ontology at the application ontology based on the actual situation and the available data that present individuals of this concepts. For example, in the case of an SO-DSPL that needs service configurations(i,e amazon service), user skills profile is important however in the case of an SO-DSPL of e-commerce, the user skills profile is not important to be consideration in recommendation process.

4.2.1 OfSoPLR application user dimension sub-ontology

At domain user dimension sub-ontology, all relative information is modeled however, no instances are created.At application level, a configuration will be recommended according to user needs and his contextual information, thus, to perform this contextual process application user dimension sub-ontology is created to conceptualize current user contextual information,as shown in Fig.8. This sub-ontology addresses the competency questions CQ2 and CQ3. All kind of current user context element are derived from core user context element presented at domain user sub-ontology.A current user has a current user skills profile that is derived from user skills profile, a current user interest profile that is derived from user interests profile, a current user personal information that is derived from user personal information and current user preference profile that is derived from user preference profile. the relation of derivation designates the selection of the concepts to be used from the domain ontology at the application ontology. For example, user preference profile at domain engineering contains all possible information related to the preference topic, however, in application engineering, not all information will be available and the needed preference's information differs from an SPL to another and from a situation to another. All current user context element all characterized by a time that refers to the date time of the monitored current user context and a duration that to the period in which this information is valid, for example user personal information profile contains the attributes first name, last name, age and email which are valid for all times, however the location of a user is variable so there duration can be expressed by a number .

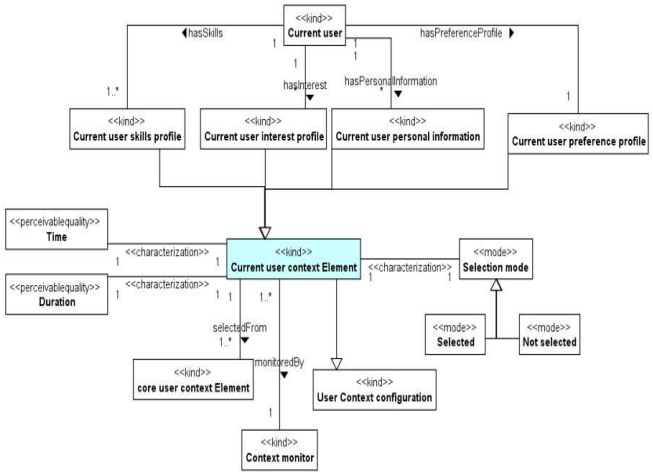


Fig. 8. OfSoPLR application user dimension sub-ontology

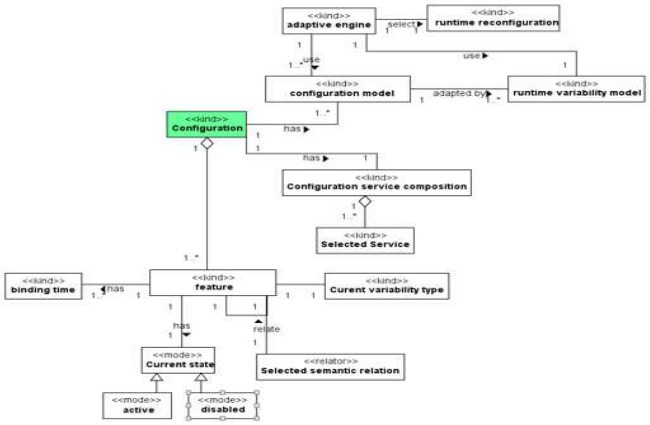


Fig. 9. OfSoPLR application configuration dimension sub-ontology

4.2.2 OfSoPLR application configuration dimension sub-ontology

In the application level, based on current user context and needs, a selection of features that fulfill these requirement is defined based on the SO-DSPL design model and a configuration is created from this selection of features. Thus, from the domain SO-DSPL sub-ontology, a configuration sub-ontology is derived to conceptualize configuration knowledge based on the product line ones and it addresses the competency questions CQ2 and CQ3.As shown in Fig.9, a configuration is composed of features. According to the framework of the dynamic software product lines, a feature can be activated or disactivated in order to perform dynamic adaptation, this characteristic is represented in our presented sub-ontology by the mode "current state" that can be divided into active and disabled mode. Each two features are related by a semantic relation derived from SO-DSPL sub-ontology in order to infer new knowledge or deduce new semantic relations according to the current contexts. A configuration has one or more configuration model that captures the possible states of the features with respect to the FM restrictions. The configuration model is adapted by a runtime variability model, which automates the mechanism of activating, disactivating, adding or removing features in the configuration model. The configuration model and the runtime variability mode are used by the adaptive engine in order to selected the best runtime reconfiguration based on the current configuration context.

4.2.3 OfSoPLR application service dimension sub-ontology

Likely to the other application sub-ontology, the application service sub-ontology is derived from the domain service ontology to address the competency questions C2 and CQ3,as shown Fig.10. Indeed, the selected service is the service which is mapped or used by a selected feature in the current configuration. The selected service is selected from the service set created by the domain service sub-ontology and inherits all knowledge according to the core service such as service profile, service grounding, service model, etc. the selected service by the current configuration are structured into a service composition in order to define their order and their combination way. At an instant t, a selected service has a service instance, which has a current service context.The current service context element, in turn, is composed by QOS defined by the domain application service subontology. The QOS attributes are monitored by a service monitor in order to get their value in a specific time. Also, a selected service is characterized by a critical level which is measured or deduced according to the current context.

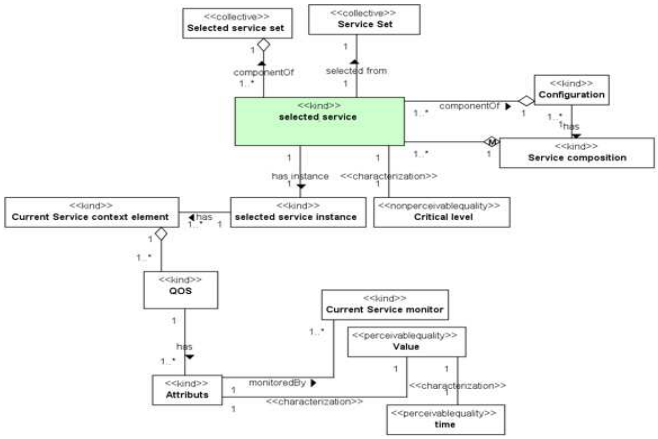


Fig. 10. OfSoPLR application service dimension sub-ontology

4.2.4 Application recommender dimension meta-model

At application engineering, a recommendation will be performed for a particular user in a defined context .To conceptualize knowledge used by the recommender engine at this level, an application recommender dimension meta-model is defined to adress the competency questions CQ2,CQ3,CQ4, CQ5,CQ7 and CQ8, as shown in Fig.11.The configuration recommender engine captures the actual configuration context element such as current user context element and use the module Situation analyzer to decide which recommendation method will be used for the current context. Thus, this module uses previous decisions produced at previous recommendation or at domain level, it reuse also models, initial recommendations and previous

recommendation from the OfSoPLR domain recommender sub-ontology. After produced the actual decision, a recommendation is produced based-on. In this way, all the processing linked to a recommendation process is developed at the domain level to prepare recommendation assets which will be reused by the recommendation system at application level. Consequently, real-time recommendation will be easier and more efficient.

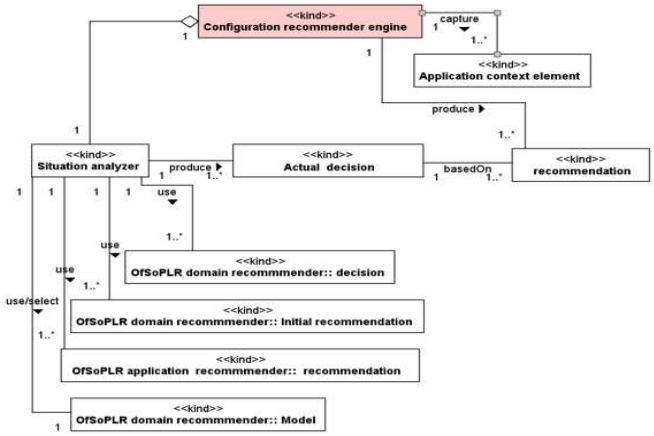


Fig. 11. Application recommender dimension meta-model

5 An OWL Operational Version of OfSoPLR

OfSoPLR ontology is to be used in a SO-DSPL recommender system context in order to conceptualize and deduce knowledge to perform the most suitable recommendation . Based on the proposed ontoUML meta-models, an operational version must be designed and implemented. The same ontology can be used to produce a number of different operational versions, each one considering a target language/environment.Operational ontologies are not focused on representation adequacy, but are designed with the focus on guaranteeing desirable computational properties. A design phase, thus, is necessary to bridge the gap between the conceptual modeling of the ontology and it's coding in terms of a specific operational ontology language (such as, for instance, OWL, RDFS, F-Logics).We chose OWL as target operational language, since it is the most used language in the scenarios we intend to use this operational version.For implementing OWL-OfSoPLR, we follow the transformation rules from OntoUML to OWL proposed in [22]. These rules guide the transformation of OntoUML concepts and relations to OWL classes and properties.

Acording to the OfSoPLR domain service dimension sub-ontology, Core Service is a superclass of Atomic and Composite . Concepts of the type kind are mapped as disjoint subclasses. The concepts of the type subkind are also mapped as subclasses of their respective superclasses. For example, Core Service is a kind and has as subkind types the following Core Service: Atomic and Composite. These concepts are mapped to disjoint OWL subclasses, as shown in Fig.12. In this figure, we present part of the transformation discussed above, considering Core Service and it subtypes based on the same format of transformations presented in [22].

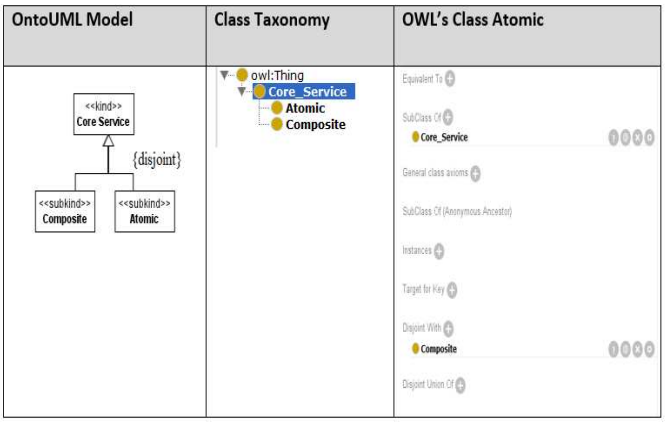


Fig. 12. Example of transformation to OWL

6 OfSoPLR Evalutation

In order to evaluate OfSoPLR,we propose an evaluation in two steps. First, we performed a verification activity by means of expert judgment, in which we checked if the concepts, relations and axioms defined in OfSoPLR are able to answer its competency questions, in an assessment by human approach to ontology evaluation. OfSoPLR was then implemented in OWL, and the resulting operational ontology (OWL-OfSoPLR) was also tested. Test cases were designed and exercised in the context of a sub-ontology, in order to check if OWL-OfSoPLR is able to answer the competency questions.

Evaluation Step 1 - Assessment by human approach to ontology evaluation

OfSoPLR evaluation started with a verification activity, when we manually checked if the concepts, relations and axioms defined in OfSoPLR are able to answer its competency questions (CQs). Table 1 illustrates an extract of this verification process, showing which elements of the ontology (concepts, relations and properties) answer the Competency Questions CQ1 and CQ2, where we present for CQ2 the part relative to the OfSoPLR domain service dimension.

Evaluation Step 2 - Ontology Testing

For each specific CQ, we developed a set of test cases, by implementing the CQs as SQWRL.For example, to answer the CQ1, the SQWRL SQ1 is created:

SQ1:

SODSPLRecommenderEngine(?s) $\wedge$ $usedBy(?y, ?s) - > sqwrl : select(?y)$

Finally, after executing a test case, we compared the returned results with the expected results to determine whether the test case passed or failed. If the results match, then OWL-OfSoPLR passed in this test case. Otherwise, we need to analyze if the problem is in the conceptual model of OfSoPLR, in its implementation (OWL-OfSoPLR), or even in the formulation/implementation of the CQs. For running the test cases, we used Protégé. As we can see by contrasting the actual result with the expected result, this test case passed.

TABLE 1
Verifying OfSoPLR concepts and relations

CQ	Concepts, Relations and Properties
CQ1	SO-DSPL is usedBy SO-DSPL recommender engine Core service is usedBy SO-DSPL recommender engine User context configuration is usedBy SO-DSPL recommender engine Core context element is usedBy SO-DSPL recommender engine Configuration is usedBy SO-DSPL recommender engine Selected Service is usedBy SO-DSPL recommender engine
CQ2	Core Service has interface provider Core Service has interface receiver Core Service has interface provider Service Instance Core Service has Service Grounging Core Service has Service Profil Core Service has Service Model Core Service has Critical level Service Profil includes QOS Service Profil has attribut Attribut has weight Core Service is executed on an environment Core Service has Service variability model Core Service has USLA model Core Service is used by SO-DSPL recommender engine Core Service set is composed of Core Service

7 Conclusion and future work

Our work provides a further contribution towards a knowledge ontology for recommender system in Service-Oriented Software Product lines.Using this ontology, relevant data for contextual recommendation are conceptualization to be used in configuration process.The proposed ontology is divided into two sub-ontologies according to domain and application engineering,and each sub-ontology is divided into three sub-ontologies in order to cover different dimensions of the conceptualized knowlegde. As future work, we plan to search for other dimensions related to the recommendation such as context of the adaptive engine and the external environment. Furthermore, we will represent real world situations with the proposed ontology and instantiate its concepts and relations using data extracted from real SO-DSPL, in a data-driven approach to ontology evaluation .

References

[1] K. POHL, G. Böckle, and F. van der Linden,*Software Product Line Engineering: Foundations, Principles and Techniques.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[2] B. LUCIANO,G. SAM ,and P. LILIANA,. *Service-Oriented Dynamic Software Product Line.* COMPUTER, 42 - 48,2012

[3] R. CAPILLA,J. BOSCH,P. TRINIDAD,A. RUIZ-CORTES,AND M. HINCHEY.*An Overview of Dynamic Software Product Line Architectures and Techniques: Observations from Research and Industry.* The Journal of Systems and Software,2014.

[4] D. Benavides,S. Segura,A. Ruiz-Cortés . *Automated analysis of feature models 20 years later: a literature review.* Inf Syst 2010;35(6):615–708 .

[5] C. Aggarwal.*Recommender Systems:The Textbook.* Springer International Publishing Switzerland ,2016.

[6] A. PALIWAL,B. SHAFIQ,J. VAIDYA,H. XIONG,and N. ADAM.*Semantics-based automated service discovery*. IEEE Transactions on Services Computing, 260–275,2012.

[7] P. RODRIGUEZ-MIER,C. PEDRINACI,M. LAMA, and M. MUCIENTES. *An integrated semantic web service discovery and composition framework.* IEEE Transactions on Services Computing, 537–550,2016.

[8] N. ZHANG,J. WANG and Y. MA. *Mining domain knowledge on service goals from textual service descriptions.* . IEEE Transactions on Services Computing, in press,2017.

[9] Z. JIEMING,H. PINJIA,Z. ZIBIN and R. MICHAEL. *Online QoS Prediction for Runtime Service Adaptation via Adaptive Matrix Factorization.* IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS,2017

[10] A. BENALIL , B. EI ASRI, H. and KRIOUILE . *Cloud Environment Assignment:A Context-aware and Dynamic Software Product Lines-Based Approach.* IEEE 2016.

[11] W. CHEN, Q. WEIWEI, Z. ZIBIN, W. XINYU, and Y. XIAOHU. *QoS Prediction of Web Services Based on Two-Phase K-Means Clustering.* IEEE International Conference on Web Services 161-168,2015.

[12] C. FUZAN ,L. MINQIANG, W. HARRIS , and X. LINGLI . *Web service discovery among large service pools utilising semantic similarity and clustering.* Enterprise Information Systems, 452-469 , 2017.

[13] R. JACKSON,S.N ALOISIO , AND C VINICIUS. *Toward a QoS Based Run-time Reconfiguration in Service-oriented Dynamic Software Product Lines.* Proceedings of the 16th International Conference on Enterprise Information Systems, 460-465 , 2014.

[14] X. RUIBIN, W. JIAN, Z. NENG,and M. YUTAO. *Deep hybrid collaborative filtering for Web service recommendation.* Expert Systems With Applications, 191–205,2018.

[15] JA. Pereira , K . Constantino and E. Figueiredo . *A systematic literature review of software product line management tools.* In: Proceedings of the International Conference on Software Reuse (ICSR). Springer; 2015. p. 73–89 .

[16] JA. Pereira ,P. Matuszyk K ,S. Krieter,M. Spiliopoulou and G. Saake.*Personalized recommender systems for product-line configuration processes* Computer Languages, Systems & Structures, 1–21, 2018.

[17] K. Skillen, L. Chen, C. Nugent, M. Donnelly, W. Burns, and I. Solheim. *Ontological user profile modeling for context-aware application personalization.* in: Proceedings International Conference on Ubiquitous Computing and Ambient Intelligence, Springer, 2012, pp. 261–268.

[18] L. Zhong-Jun, L. Guan-Yu and P. Ying. *A method of meta-context ontology modeling and uncertainty reasoning in SWoT.* Proceedings of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (2016) 128–135.

[19] Web ontology language for services, W3C member submission 2004. http://www.w3.org/Submission/OWL-S/width=3.5in, height=3.5in

[20] G. Lucia, GI. LUISA and R. MAZO. *IDENTIFYING DEAD FEATURES AND THEIR CAUSES IN PRODUCT LINE MODELS: AN ONTOLOGICAL APPROACH* Dyna (Medellin, Colombia) 81(183):68-77,2014

[21] G. Guizzardi. *Ontological foundations for structural conceptual models.* The Netherlands: Universal Press, ISBN 90-75176-81-3,2005

[22] P. Paulo F. Barcelos, V. Amorim dos Santos, F. Brasileiro Silva, M. Monteiro and A. Salles Garcia.*An automated transformation from OntoUML to OWL and SWRL* The Ontology Research Seminar in Brazil (ONTOBRAS), Belo Horizonte, MG, 130-141,2013

ADETermino: a Platform to Construct Adverse Drug Events Corpus from MEDLINE Towards ADE Ontology Enrichment

NOUIRA Kaouther
B.E.S.T.M.O.D
ISG, Université de Tunis
Kaouther.nouira@gmail.com

TEBBEB Sirine
B.E.S.T.M.O.D
ISG, Université de Tunis
sirine-tebbeb@hotmail.fr

Abstract

This paper presents a new Web platform (ADETermino) which allows the automatic construction of an Adverse Drug Events (ADE) corpus starting from the free search engine PubMed which accesses to the MEDLINE database of references and abstracts of articles from journals covering medicine, nursing, pharmacy, dentistry, veterinary medicine, healthcare, biology and biochemistry [1][6]. The aim of this platform is to facilitate semi-automated enrichment analysis for ADE Ontology [2] [3]. It offers an automatically updated ADE Corpus.

Keywords: Adverse Drug Event, Corpus, Ontology, MEDLINE, Reference Management.

1 Introduction

Adverse Drug Event (ADE) is an incident which occurs throughout the healthcare system when the medical intervention is related to drugs [4] (i.e. administration of Panadol-acetaminophen to a patient with severe liver injury includes acute liver failure resulting in liver transplant or death, administration of Panadol-acetaminophen to patient using Teriflunomide may cause liver problems). In order to propose an ADE preventive system, Nakhla and Nouira in [5] suggest to construct an ADE ontology. Their target was to standardize the knowledge representation of the ADE domain and to allow their system using standard structured data.

The weakness of this ontology is the fact that it is static. So this paper proposes a new approach to enrich the ontology and enables it to change whenever a new publication concerning ADEs appears in MEDLINE database.
This paper is organized as follows: Section 2 presents the requirements Definition. Then Section 3 discusses steps to construct an ADE corpus. After that, Section 4 discusses results. Finally Section 5 presents the conclusion and the future works.

2 Requirements Definition

The aim of this work is to build a tool so called ADETermino, able:

- to connect to MEDLINE database,
- to import, automatically, the papers dealing with ADE into a specific database so called "ADE Corpus" whenever there is a new publication.

Several free and paid tools are available. Such tools are called Reference Management (RM) tools. RM tools are widely used to store, organize and manage bibliographic references for research papers [8]. There are many RM tools able to import files from MEDLINE (i.e. Bibdesk, EndNote, RefDB, refbase, Zotero, colwiz). All these tools are oriented medical domain in general and not specialized on ADE field. All these tools need human intervention to import papers [6]. All these tools make redundancy, every time when they are sought to access the database. Gilmour et al. [7] can download former files. To address these problems we propose:

- To standardize the file structure in order to reduce their spacial complexity,
- To automate the process of search,
- To decrease the time duration.

3 Steps to construct ADE Corpus

This subsection presents five steps to construct an ADE specialized corpus as illustrated in Figure 1:

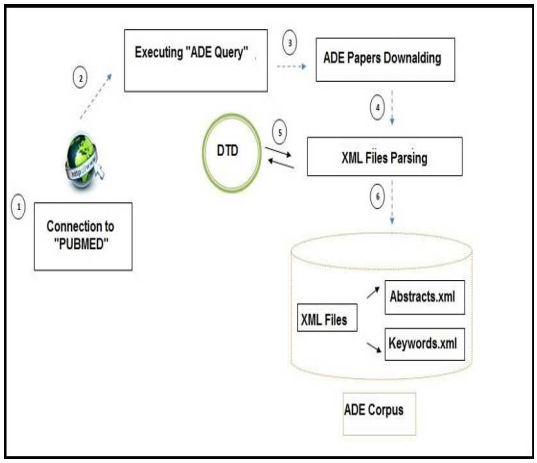
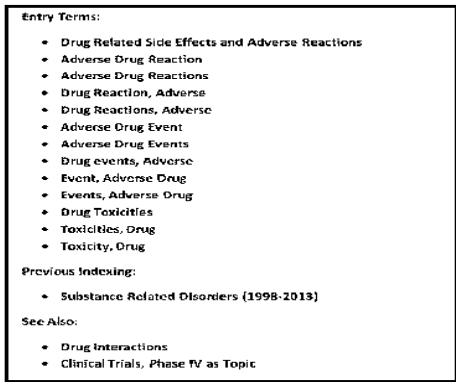


Figure 1: Automatic connection to MEDLINE.

Step 1: Automatic connection to MEDLINE ADETermino automatically connects to MEDLINE via PubMed interface [8]. It uses Medical Subject Headings (MeSH) vocabulary thesaurus [1] which is controlled by U. S. National Library of Medicine (NLM) [10]. MeSH contains over than 20.000 terms. The interest is to avoid synonymy and facilitate database querying [9] [10].

Step 2: ADEQuery (Query to extract papers dealing with ADEs).
To formulate ADEQuery, a specific ADE field keywords are selected from MeSH thesaurus (see Figure 2).
The query structure is as shown in Figure 3:


Entry Terms:

- Drug Related Side Effects and Adverse Reactions
- Adverse Drug Reaction
- Adverse Drug Reactions
- Drug Reaction, Adverse
- Drug Reactions, Adverse
- Adverse Drug Event
- Adverse Drug Events
- Drug events, Adverse
- Event, Adverse Drug
- Events, Adverse Drug
- Drug Toxicities
- Toxicities, Drug
- Toxicity, Drug

Previous Indexing:

- Substance Related Disorders (1998-2013)

See Also:

- Drug Interactions
- Clinical Trials, Phase IV as Topic

Figure 2: keywords selected form MeSH Thesaurus

Step 3: ADE papers Downloading IDLIST is the result of ADEQuery. This file contains identifiers (PMID: PubMed Identifier) of all

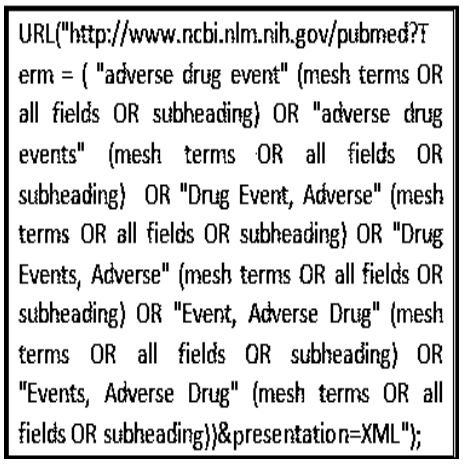

URL("http://www.ncbi.nlm.nih.gov/pubmed?Term = ("adverse drug event" (mesh terms OR all fields OR subheading) OR "adverse drug events" (mesh terms OR all fields OR subheading) OR "Drug Event, Adverse" (mesh terms OR all fields OR subheading) OR "Drug Events, Adverse" (mesh terms OR all fields OR subheading) OR "Event, Adverse Drug" (mesh terms OR all fields OR subheading) OR "Events, Adverse Drug" (mesh terms OR all fields OR subheading))&presentation=XML");

Figure 3: ADEQuery

papers selected by the ADEQuery (see Figure 4).
The advantage of the use of IDLIST is the fact

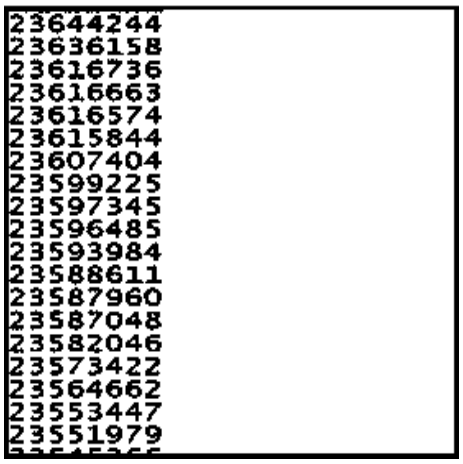

23644244
23636158
23616736
23616663
23616574
23615844
23607404
23599225
23597345
23596485
23593984
23588611
23587960
23587048
23582046
23573422
23564662
23553447
23551979

Figure 4: IDLIST

that it saves the old papers already selected by ADETermino and whenever it is restarted it will download only the recent ones. ADETermino will use IDLIST to download papers as "WEB format" and convert them to "XML format".

Step 4: XML files parsing the step 3 resulting XML file contains a lot of unused tags which cause a slowdown in the concept parse process. ADETermino normalize the structure of the XML paper and store it as an XML file. This normalization is based on a Document Type Definition (DTD) [11] proposed in order to describe the new XML file model and to know which tags must be selected

from the initial XML file (see Figure 5).

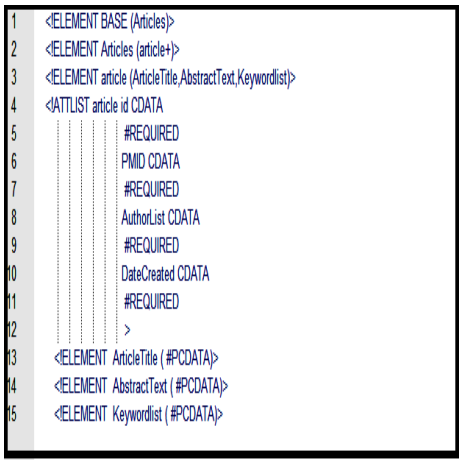


Figure 5: DTD for initial XML files

Step 5: ADE Corpus creation After XML parsing according to the proposed DTD, the new file will be stored in a corpus called ADE Corpus. An ExteractionTag Algorithm will be executer to split the XML file into two files "Abstracts.xml" and "Keywords.xml". The first file contains all the abstracts selected from the XML file and the second file contains the keywords without redundancy and sorted according to the number of occurrences of each keyword (see Figure 6).

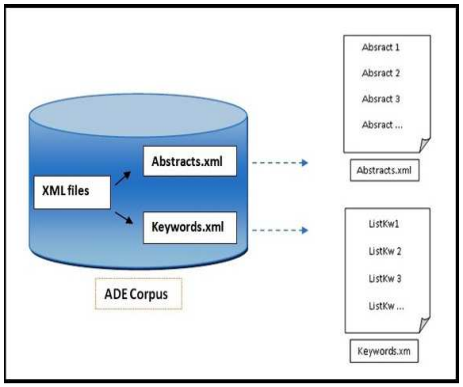


Figure 6: ADE Corpus

4 Discussion

This section presents ADETermino results: Table 1 shows the results of a comparative study between ADETermino and some popular tools.

Table 1: Comparative between ADETermino and some RM tools.

	Search Process	Average Paper Size
ADETermino	Automatic	543.4 Byte (0.44 KiloByte)
Bibdesk	Manual	3791.7 Byte (3.7KiloByte)
EndNote	Manual	3791.7 Byte (3.7KiloByte)
Zotero	Manual	3791.7 Byte (3.7KiloByte)
RefDB	Manual	3791.7 Byte (3.7KiloByte)

ADETermino outperforms the other tools. It has an automatic search process and it decrease the size of papers downloaded from MEDLINE.
For example, the average paper size of 20 papers downloaded by the different tools is $0.44KiloByte$ for ADETermino and $3.7KiloByte$ for all others. Whenever ADETermino is executed, it downloads the new papers dealing with ADE. Until now, ADE corpus contains 1685 papers from which are generated 1685 abstracts and 362 keywords without redundancy (see Table 2).

Features of ADETermino **ADETermino** :

- **Standard** : All papers have the same structure.
- **Complete and Precise** : It contains all papers found in MEDLINE specialized on ADE field.
- **Up-to-date** : The latest papers added recently in MEDLINE are automatically added.
- **Coherent and non-redundant** : Don't find the same paper more than once.

Table 2: ADE Corpus Content

	Abstracts Number	Keywords Number
ADETermino	1685	362

5 Conclusion

This paper proposes a new web platform so called **ADETermino**.
ADETermino uses MEDLINE database to construct and update automatically an ADE Corpus. ADE Corpus will be used to enrich the ADE Ontology whenever a new publication concerning ADE appears in MEDLINE database,on the first

hand, and it will be useful to researchers and experts to find excatly that they are looking for on ADE field on the second hand.
The obtained results show the performance and the effectiveness of ADETermino compared to other RM tools.

- ADETermino standardizes the file structure in order to reduce space complexity.
- ADETermino automate the search process.
- ADETermino decreases the downloaded paper size in order to reduce time complexity.

Future works will be fulfilled :

- To access to different bibliographic databases sources such as Science Direct, BIOSIS Citation Index, Current Contents ans Science Citation Index...
- To extract concepts, relationships and instances from ADE Corpus to enrich ADE Ontology.

References

[1] Mouillet, E. (2008). PubMed 2009 La nouvelle interface de recherche avancee (Advanced Search). Cahiers de Santé, 18(4).

[2] Agirre, E., Ansa, O., Hovy, E., and Martnez, D. (2000). Enriching very large ontologies using the WWW. arXiv preprint cs/0010026.

[3] Mrabet, Y., Bennacer, N., and Pernelle, N. (2012). Enrichissement contrĺ de bases de connaissances ` partir de documents semi-structurś annotś. Actes des 23es Journés Francophones d'Ingńierie des Connaissances-IC 2012, (ISBN 978-2-7466-4577-6), 17-32.

[4] Nebeker, J. R., Barach, P., and Samore, M. H. (2004). Clarifying adverse drug events: a clinician's guide to terminology, documentation, and reporting. Annals of internal medicine, 140(10), 795-801.

[5] Nouira, K., and Nakhla, Z.(2012). Development of Ontology for the Representation of Adverse Drug Events of Diabetes Disease. International Journal of Computer Applications, 42.

[6] Hardy, G. J., and Robinson, J. S. (1996). Subject guide to US government reference sources. Libraries Unlimited.

[7] Gilmour, R., and Cobus-Kuo, L. (2011). Reference management software: a comparative analysis of four products. Issues in Science and Technology Librarianship, 66(66), 63-75.

[8] Emanuel, J. (2013). Users and citation management tools: use and support. Reference Services Review, 41(4), 639-659.

[9] Meusel, R., Niepert, M., Eckert, K., and Stuckenschmidt, H. (2010). Thesaurus extension using web search engines. In The Role of Digital Libraries in a Time of Global Change (pp. 198-207). Springer Berlin Heidelberg.

[10] Boudry, C., and Bozet, G. (2004). Recherche bibliographique en biologie et en mdecine: du bon usage de Medline Pubmed. M/S: mdecine sciences, 20(8-9), 804-807.

[11] Bosak, J., Bray, T., Connolly, D., Maler, E., Nicol, G., Sperberg-McQueen, C. M., ... and Clark, J. (1998). W3c xml specification dtd.

An Application of the Peak Over Threshold Method for Ozone Data Modeling

NOUIRA Kaouther
BESTMOD Laboratory, ISGT
Université de tunis
Tunis, Tunisia
https://orcid.org/0000-0002-9001-3686

AYARI Samia
BESTMOD Laboratory, ISGT
Université de tunis
Tunis, Tunisia
https://orcid.org/0000-0003-3894-1650

CHETOUANE Ines
BESTMOD Laboratory, ISGT
Université de tunis
Tunis, Tunisia
chetouaneines@yahoo.fr

***Abstract*— This paper presents a new approach to monitor air quality. It studies the statistical behavior of the extreme Ozone concentrations by fitting the parameters of the Generalized Pareto Distribution. The results show the good performance of this method in estimating the return level of Ozone concentrations for different return periods.**

Keywords*— Extreme Values Theory, Peak Over Threshold method, Generalized Pareto Distribution, threshold selection, Ozone

I. INTRODUCTION

Several pollutants are responsible for air pollution. Ozone is considered among the most air pollutant that have harmful effects for environment and human health [1]. In fact, Ozone can damage ecological resources [2] and can cause complex health problems such as severe cardiovascular and lung diseases [3].

Many studies were done to monitor Ozone concentration [4], [5], [6]. Most of them are limited to forecast the future concentration value but never to know the probability of occurrence of future extreme Ozone concentrations. Such problem can be solved with the Extreme Values Theory (EVT).

EVT is a discipline that seeks to model and analyze the stochastic behavior of extreme values [7]. In recent years, EVT is considered as a field of research in several areas [8] such as financial field [9], demography field [10], environment field [11] … etc. It has been applied in order to estimate the extreme events of phenomena.

EVT proposes two approaches for modeling extreme events, the Block Maxima (BM) approach [12] and the Peak Over Threshold (POT) approach [13]. Such approaches are used to estimate the probability of occurrence of extreme events. Each approach has its own distribution. The Generalized Extreme Values Distribution is used in BM approach (to analyze blocks maxima) [14] and the Generalized Pareto Distribution is used in the POT approach (to analyze the threshold exceeding) [15].

The BM approach is a parametric approach that consists of dividing the sample of observations into equally sized blocks and calculating the maximum of each block [16]. BM approach requires the estimation of the Generalized Extreme Values Distribution parameters to analyze extreme events. But, the fact of dividing the sample into a set of blocks and taking the maximum value of each block may cause information loss, because the block can contain several maximum values, as it may contain no maximum value [17].

However, the POT approach improves the performance by analyzing and modeling all observations that exceed a specific threshold using the Generalized Pareto Distribution (GPD) [22]. For this reason the POT approach will be used in this paper to estimate extreme return levels of Ozone concentrations.

The paper is organized as follows: the second section presents the principle of the POT model and the GPD parameters by using the maximum likelihood estimator, the third section discusses the graphical approaches to selecting the best threshold, section 4 describes the parameters estimation of the GPD, section 5 deals with the model validation that is based on the calculation of the diagnostic plots and finally the section 6 presents conclusions and perspectives.

II. THEORETICAL DISCUSSION

The POT approach is used to extract from a sample of observations a set of extreme values.

It consists of:

Choosing a threshold u.

Building a sample of all observations that exceed the threshold (Extreme Values).

Analyzing and modeling all observations that exceed a specified threshold using the GPD.

The threshold value should not be very small to avoid regular observations, and should not be very large to miss extreme values. The choice of the threshold level can be done through the GPD mean excess function which offers a graphical method for selecting the appropriate threshold [18]. It is interested in approaching the asymptotic distribution of a random variable x as long as it exceeds the threshold.

The difference between the observed value xj and the threshold u is the excess calculated in eq.1.

$$y_j = x_j - u. \tag{1}$$

The probability that a random variable x of the sample exceeds a specified high threshold level is expressed by the unknown distribution function F. This function can be estimated by a conditional excess distribution function Fu(y) defined as.

$$F_u(y) = P(x - u < y \mid x > u) = \frac{F(u+y) - F(u)}{1 - F(u)}. \tag{2}$$

For $0 \leq y \leq x_F$, where x_F is either a finite or infinite right endpoint of the underlying distribution F.

The GPD of observations above threshold is:

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi \frac{x}{\beta})^{\frac{1}{\xi}}, \xi \neq 0 \\ 1 - e^{-\frac{x}{\beta}}, \xi = 0 \end{cases}, \quad (3)$$

whith:

$$\begin{cases} x \geq 0, if\ \xi \geq 0 \\ 0 \leq x \leq -\frac{\beta}{\xi}, if\ \xi < 0 \end{cases}.$$

The Generalized Pareto Distribution has two parameters:

$\beta > 0$: Scale parameter representing the variance.

$\xi \in \mathbb{R}$: Shape parameter representing the tail index or the extreme value index.

which are estimated by the maximum likelihood [19].

III. EMPIRICAL DISCUSSION

The POT approach is applied on an hourly series of Ozone concentration in Gabes region in Tunisia during year 2011.

The POT method in based on the following points:

- Threshold selection by graphical approaches.
- Studying the statistical behavior of excess.
- Fitting the parameters (β and ξ) of GPD.

The mean residual life plot, the modified scale and the shape plots are graphical approaches for choosing the best threshold.

According to [19] and [20], the threshold is in the interval where the mean excess function decreases linearly with the threshold values.

In the mean residual life plot (see fig.1), the mean excess function decreases linearly in the *[0,45]* interval.

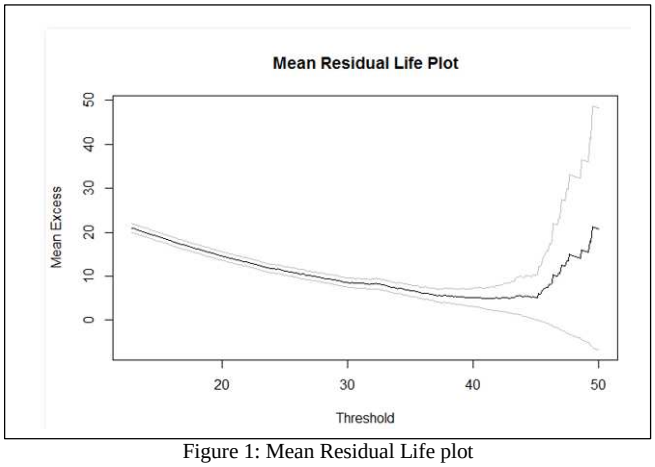


Figure 1: Mean Residual Life plot

Fig.2 and fig.3 present the second graphical approach for threshold selection which says that the adequte threshold corresponds to constant modified scale and shape parameters [20] which gives the same interval.

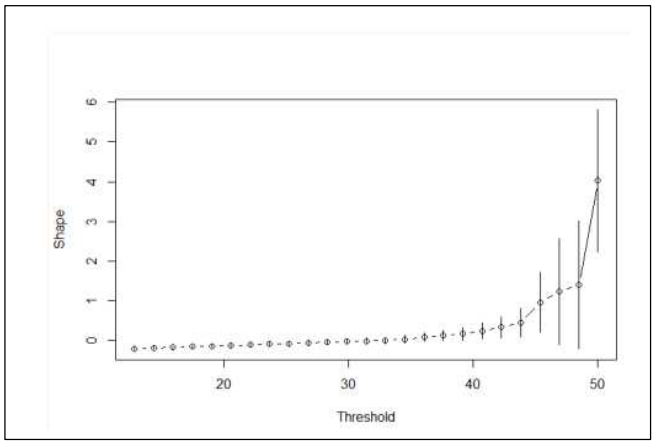


Figure 2: Shape parameter plot

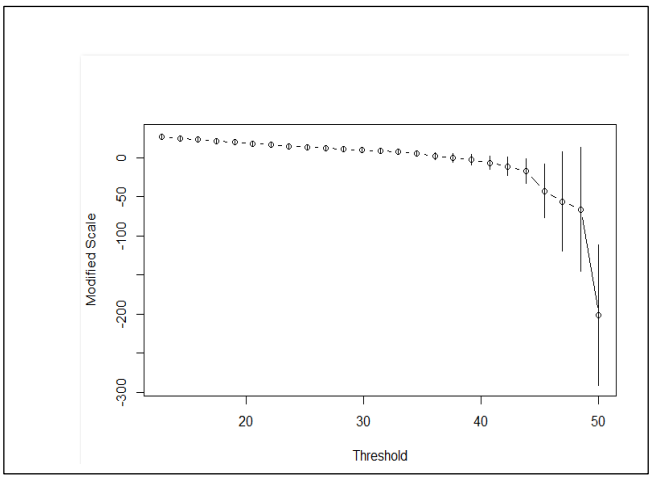


Figure 3: Modified scale parameter plot

After determining the threshold interval, a simple estimate of the distribution parameters for different threshold values is summerized in table 1.

TABLE I. ESTIMATION OF SCALE AND SHAPE PARAMETERS

u	10	20	30	35
Scale (β)	27.4873 (1.5037)	16.0905 (0.9227)	8.69476 (0.64127)	6.37417 (0.5921)
Shape (ξ)	-0.2281 (0.0167)	-0.1259 (0.0180)	-0.01587 (0.03305)	0.04838 (0.0468)
Loglikelihood	2851.884	-1212.57	-723.7736	-449.5943
Number of exceedances	349	332	14	155
U	40	41	42	45
Scale (β)	3.9503 (0.5575)	3.2761 (0.5228)	3.0819 (0.5583)	1.2414 (0.4006)
Shape (ξ)	0.2045 (0.0915)	0.2765 (0.1132)	0.3235 (0.1318)	0.8366 (0.3078)
Loglikelihood	-213.9975	-179.8073	-142.0439	-65.69194
Number of exceedances	83	73	58	32

In table 1 scale and shape parameters of the GPD are estimated for different threshold values in the interval *[0,45]*.

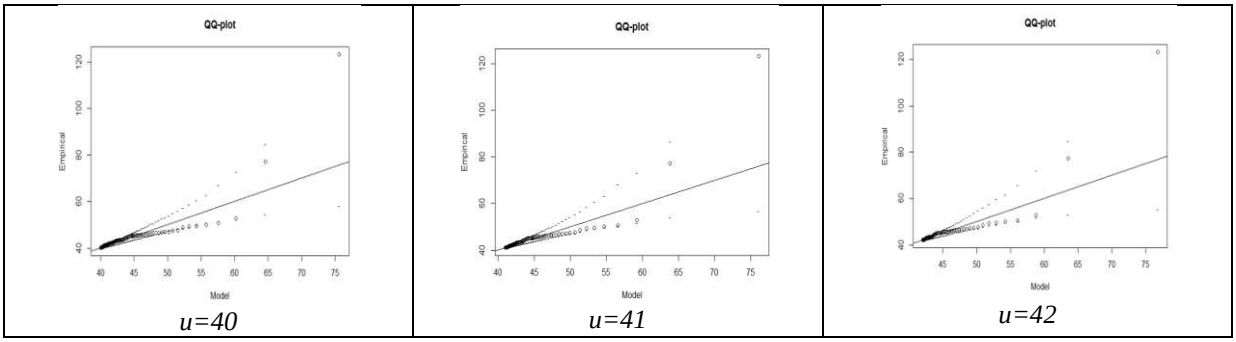
The best threshold is selected when the POT parameters are approximately stable [21].

For the values of thresholds *40*, *41* and *42*, the shape and the scale parameters are roughly constants. So, we conclude that the threshold is optimal for these several values.

QQ-pot (see fig.4), probability plot (see fig.5), density plot (see fig.6) and return level plot (see fig.7) are used as diagnostic plots for threshold values in order to assess and validate the quality of a fitted GPD. Diagnostic plots represent a validation model used to verify the quality of the estimated model. Data is well fitted by the GPD if both QQ-plot (see fig. 4), probability plot (see fig. 5) are approximated by a linear line. This improves that the data is well fitted by the GPD. The diagnostic plots for each of these three possible threshold values gave the same pattern (see fig. 4, 5, 6 and 7).

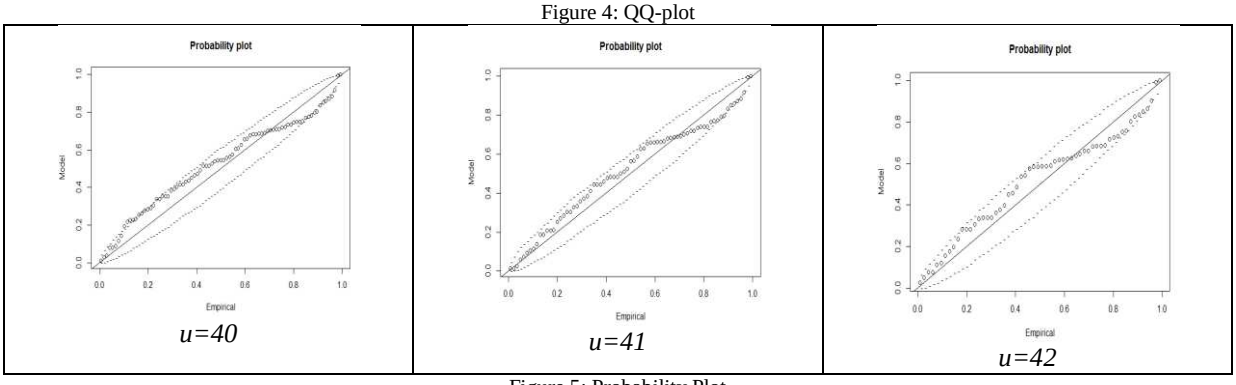
Therefore, the model can be applied to estimate the return level of Ozone concentrations for different return periods. Refer to fig.7 the Ozone concentration will have a peak after a year (in 2012) with an estimated value of 43, a second peak (in 2013) with an estimated value of 44, a third peak (in 2016) with an estimated value of 46 and so on.

with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".



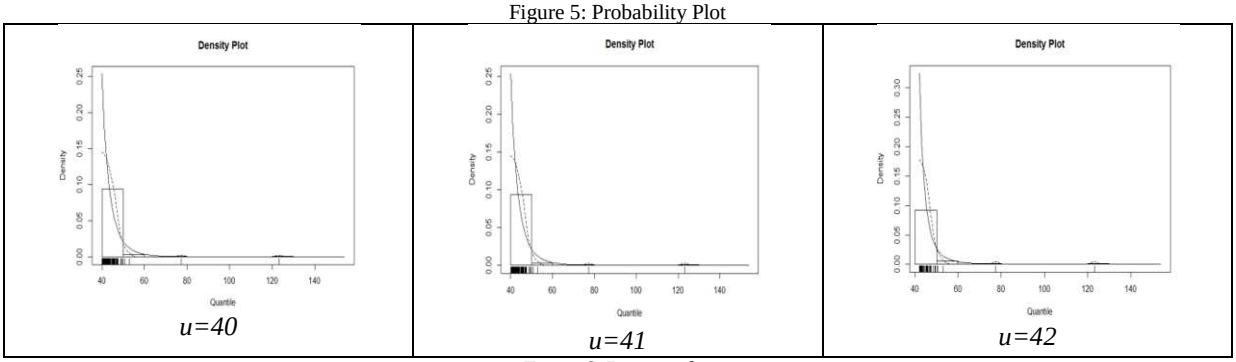
u=40 *u=41* *u=42*

Figure 4: QQ-plot



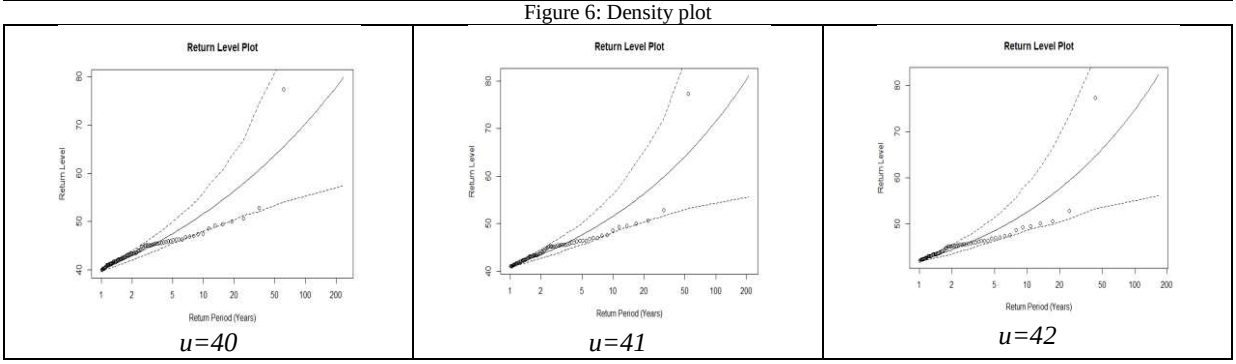
u=40 *u=41* *u=42*

Figure 5: Probability Plot



u=40 *u=41* *u=42*

Figure 6: Density plot



u=40 *u=41* *u=42*

Figure 7: Return Level plot

IV. Conclusion

Several deterministic and statistical models have been developed to evaluate extreme values and to estimate the probability of occurrence of rare events.

In this context, the EVT is used to predict extreme pollutants concentrations. The POT approach using GPD is used to estimate extreme return levels of Ozone concentrations (an hourly series of Ozone concentration in Gabes region in Tunisia during year 2011 is used) for several return periods.

The results show that the method worked well by giving good estimates of extreme values for 200 years to come. The estimated values of the first two years (2012 and 2013) have been verified and validated. Such results motivate the use of POT approach in anomaly detection in medical field to estimate the probability of occurrence of a heart attack or a cerebrovascular accident.

References

[1] B. Brunekreef and S. T. Holgate. Air pollution and health. The lancet, 360(9341):1233–1242, 2002.

[2] W. H. O. R. O. for Europe and W. H. Organization. Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. World Health Organization, 2006.

[3] M. L. Bell, A. McDermott, S. L. Zeger, J. M. Samet, and F. Dominici. Ozone and short-term mortality in 95 us urban communities, 1987-2000. Jama, 292(19):2372–2378, 2004.

[4] S. Ayari, K. Nouira, and A. Trabelsi. A hybrid arima and artificial neural networks model to forecast air quality in urban areas: Case of tunisia. Advanced Materials Research, 518:2969–2979, 2012.

[5] J. Yi and V. R. Prybutok. A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area. Environmental Pollution, 92(3):349–357, 1996.

[6] A. C. Comrie. Comparing neural networks and regression models for ozone forecasting. Journal of the Air & Waste Management Association, 47(6):653–663, 1997.

[7] L. De Haan and A. Ferreira. Extreme value theory: an introduction. Springer, 2007.

[8] J. Galambos. Extreme value theory for applications. Springer, 1994.

[9] P. Embrechts, C. Klüppelberg, and T. Mikosch. Modelling extremal events: for insurance and finance, volume 33. Springer, 1997.

[10] V. Marimoutou, B. Raggad, and A. Trabelsi. Extreme value theory and value at risk: application to oil market. Energy Economics, 31(4):519–530, 2009.

[11] S. G. Coles and D. Walshaw. Directional modelling of extreme wind speeds. Applied Statistics, pages 139–157, 1994.

[12] M. Rocco. Extreme value theory in finance: A survey. Journal of Economic Surveys, 28(1):82–108, 2014.

[13] V. P. Fernández. Extreme value theory: value at risk and returns dependence around the world. Technical report, Centro de Economía Aplicada, Universidad de Chile, 2003.

[14] S. Kotz and S. Nadarajah. Extreme value distributions. World Scientific, 2000.

[15] V. O. Andreev, O. B. Okunev, and S. E. Tinyakov. Extreme value theory and peaks over threshold model: An application to the russian stock market. New York Science Journal, 3(6):102–107, 2010.

[16] M. Rocco. Extreme value theory in finance: A survey. Journal of Economic Surveys, 28(1):82–108, 2014.

[17] T. Roncalli. Théorie des valeurs extrêmes-modélisation des evenements rares pour la gestion des risques. Groupe de Recherche Opérationnelle, Crédit Lyonnais, 2002.

[18] V. O. Andreev, O. B. Okunev, and S. E. Tinyakov. Extreme value theory and peaks over threshold model: An application to the russian stock market. New York Science Journal, 3(6):102–107, 2010.

[19] D. E. Allen, A. K. Singh, and R. J. Powell. Extreme market risk-an extreme value theory approach. 2011.

[20] S. Mkhandi, A. Opere, and P. Willems. Comparison between annual maximum and peaks over threshold models for flood frequency prediction. In International conference of UNESCO Flanders FIT FRIEND/Nile project–towards a better cooperation, Sharm-El-Sheikh, Egypt, CD-ROM Proceedings, 2005.

[21] H. Rieder, J. Staehelin, J. Maeder, T. Peter, M. Ribatet, A. Davison, R. Stübi, P. Weihs, and F. Holawe. Extreme events in total ozone over arosa–part 1: Application of extreme value theory. Atmospheric Chemistry and Physics, 10(20):10021–10031, 2010.

[22] Balkema, A.A., and de Haan, L. (1974). Residual Life Time at Great Age, The Annals of Probability 2, 792-804.

A memetic algorithm for the tourist trip design with clustered POIs

Dorra Jriji
Université de Tunis,
Institut Supérieur de Gestion, LARODEC Laboratory,
Tunisia
Email: dorrajriji@gmail.com

Saoussen Krichen
Université de Tunis,
Institut Supérieur de Gestion, LARODEC Laboratory,
Tunisia
Email: saoussen.krichen@gmail.com

***Abstract*—Over the last years, intelligent applications have become available to tourists. These applications offer the possibility of recommending a list of points of interest and generating personalised routes.**

This paper deals with the Tourist Trip Design Problem(TTDP) with Clustered Points of Interest and taking into account the tourist budget, where the points of interest are grouped in clusters representing different categories. The TTDP with Clustered Points of Interest, known to be a hard constrained optimization problem and is modeled as a Team Orienteering Problem With Time Windows (TOPTW) which is a NP-hard problem. As it is a case, we propose a methaheuristic approach to generate satisfactory scenarios for TOPTW.

This paper introduces a new solution method for solving the TOP with hard Time Windows constraints. An integer programming model is formulated with the objective of finding the maximum of the sum of the collected profit by given a set of locations, each have a score, a service time, a time window and belongs to a category. The main contribution of this paper is a Memetic algorithm improved with a tabu search procedure to solve the TOPTW. The proposed algorithm is applied to benchmark instances. The results provided by the proposed algorithm are compared with those obtained by solving the mixed integer linear programming formulation. Computational results carried out on real and real-like instances show the proposed heuristic is competitive with other solution approaches in the literature.

***Index Terms*—Team Orienteering Problem, Time window, Tourist trip design problem, Clustered point of interests,heuristic optimization, Memetic algorithm, Tabu search procedure.**

1. Introduction

Tourists visiting countries and cities, for a given period of time, are faced with the challenge of choosing points of interest (POIs) would be more interesting to visit. Therefore, it is not possible to visit everything they are interested in, they have to choose what they believe to be the most interesting attractions as well as in the order they should be visited.

The challenge is to make a feasible route in order to visit the most valuable attractions with taking into account their budget and the available time span. Assuming that tourists appreciate the assistance of a personalized tourist guide when planning their trip.

We designed the problem as a Tourist Trip Design Problem (TTDP), in which tourist can find the the optimal route to their trip that involves a number of constraints such as the visiting time required for each POI, the POIs visiting days/hours, the traveling distance among POIs, the time available for sightseeing on a daily basis and the degree of satisfaction (profit) associated with the visit to each POI (based on personal profile and preferences).

The TTDP presents a trip solution that maximizes the satisfaction of the tourist while taking the above constraints. Typically, the team orienteering problem with time windows (TOPTW) is a simplified version of the TTDP [6].

In this paper, we present the Tourist Trip Design Problem with grouped POIs.We consider that the set of POIs are grouped by categories representing different types of visiting sites (culture, adventure, family,sea, restaurant, etc.). We give a set of geographical locations, each location with a service time, a score and a time window. A service cost is associated for each visited location. The main aim is to maximize the sum of the collected scores (maximum of profit for the tourist) by a fixed number of routes (one for each day of staying) while respecting the budget of the tourist.

The routes must start and end at a given starting location (which represent the hotel) and the duration of each route. As well as, a route visits locations at the appropriate time and it is limited by a length. The routing inside this problem has to be calculated in real-time in order to react to tourist actions and preferences or unexpected events.

The TTDP has become a very interesting research topic in recent years due to the high tourists re-quests and expectations. The involved challenge is to plan varied routes for tourists in order to exploit the time allocation for all days tour in the most effective way visiting the most interesting points.

Souffriau et al. [12] use a Guided Local Search to solve the simplest form of the tourist trip design problem, wherein a plan has to be devised by only taking into account the maximum allowed distance. Souffriauet al. [13] improve their approach by taking into account the opening hours of POIs for one day, then test the feasibility of the planning tour algorithm on a mobile with limited computational resources.

Rodriguez et al. [14] developed a tool involving an interactive multi-criteria techniques for providing tourists with the tour route best suited to their needs. They have applied a Tabu Search procedure for solving multi-objective metaheuristic using an adaptative memory procedure.

Tsai and Chung [8] propose a recommended system which provided personalized visiting routes for visitors in theme parks, with consideration for visitor behavior and real-time information.

Liu et al. [7] introduced a novel personalized route recommendation system to reduces total visiting time and to provide self-drive the tourists with real-time.The provided route recommendation services for self-drive tourists based on historical and real-time traffic information.

All of the above studies broadened the options of a personalized service and have focused on the functionality of tourist recommendation systems while providing specific applications. The used algorithms to solve the TTDP were primarily for expanding the functionality of systems and improve the performance. Numerous researchers pointed out that the TTDP can be modeled by using an OP.

The team orienteering problem (TOP) is known as an NP-hard combinatorial optimization problem [15]. It has been applied to several real world applications, such as the sport game of team orienteering, the recruiting of college football players, the scheduling of routing technicians service and the pickup and delivery services involving the use of common or private carriers [16].

There are various variants of the TOP exist and one of them takes Time Window constraints into account (TOPTW). Many heuristic and meta-heuristic approaches for the TOPTW are developed.

Vansteenwegen and Oudheusden [1] extends the TOP by adding the constraint of limited time availability of serviced nodes, which is the opening and closing hours of a POIs. Exact solutions are feasible for the TOPTW with very restricted number of nodes (e.g. Li and Hu [17] used on networks of up to 30 nodes).

In the light of the complexity of this problem, the TOPTW literature exclusively involves heuristic algorithms and are meta-heuristics methods that involve.

Kantor and Rosenwein [18] were the first to work on the OPTW, first they describe a simple insertion heuristic. The location which has the highest ratio (score over insertion time) is the most inserted into the tour, while respecting any of the time windows. Second, they propose a depth-first search tree algorithm that constructs partial routes, beginning from the start location by using the insertion heuristic. Partial tours are abandoned if they are non-realisable or/and if they are not likely to yield the best total score.
Righini and Salani [19] apply a bi-directional dynamic programming to solve OPTW to until obtaining the optimum combination with a technique called "decremented state space relaxation" that serves to reduce exponentially the number of states to be explored.
Vansteenwegen et al. [2] considers the TOP with time windows in the setting of a personalized electronic tourist guides, where the problem is solved by an iterated local search approach. In the PhD thesis of Souffriau [20] an iterated local search approach for the TOPTW was discussed.

More recently, Labadi et al. [21] propose a local search algorithm based on a variable neighbourhood structure for the TOPTW. In the local search routine they tries to replace a segment of a path with other nodes that are not included in a path and that offer more profit. In order to ensure this, they relate to the TOPTW an assignment problem based on that solution, the algorithm decides which arcs to insert in the path.

An heuristic algorithm based on simulated annealing (SA) for the TOPTW was proposed by by Lin and Yu [23] . A neighbouring solution is obtained on each iteration by using one of the moves swap (insertion or inversion) with equal probability. If new solution is more profitable than the currently best found solution, it is adopted and becomes the current one, else it is accepted with the probability of decreasing with increasing profit loss. After applying a certain number of iterations, the best solution found so far is further improved by applying local search.

Hu and Lim [4] deal with an iterative three-component heuristic (I3CH) for TOPTW. To our knowledge, the I3CH offers the highest quality solutions, but with high execution time. First each route is obtained from the local search, then simulated annealing procedure is inserted into a pool of routes. Finally in the route recombination step, the disjoint routes from the pool with the highest picked total profit, hence deriving to the best quality solution.

The aim of this paper is to propose new algorithm to solve the TOPTW(or the tourist trip design with clustered POIs) which combines highly effective memetic algorithm with Tabu search procedure. The proposed algorithm has been tested in terms of various performance parameters such as the solutions quality and execution time. The algorithm is compared with the results obtained by solving the MIP formulation of the proposed problem. The performance of the solution method also has been compared against the current approaches proposed to solve this problem from the literature.

The remainder of the present paper is organized as follows: Section 2 the problem description, the main assumptions and the proposed mathematical model are presented. In Section 3, the proposed solution method is introduced. In Section 4 Extensive computational experiments are presented. Finally, Section 5 concludes the paper.

2. Description of problem and mathematical formulation

As stated above, the TTDP can be modeled as a TOPTW. The main objective of the TOPTW is to maximize tourist

satisfaction by selecting points of interest (POIs) that match tourist preferences, thereby maximizing tourist satisfaction, while taking into account a multitude of parameters and constraints.

For a TOPTW, we gives a complete undirected graph $G = N, E$, each edge t_{ij} have a positive weight that represent a travel time spend from location i to location j. The nodes $i = 0$ denotes the starting and end ending location that usually corresponds to the hotel. Each POI can be included in at most one route. The sets of POIs of each category $c \in C$ constitute a partition of the set of POIs I.

The minimum and maximum number of visited POIs of category c is bounded by given values Min_c and Max_c. The number of visits to a given cluster can be fixed by using equal bounds $Min_c = Max_c$. We can also model one side limits by choosing Min_c to 0 or Max_c to the number of POIs of category c.

For each pair of nodes, i and j , is known the travel time t_{ij}.For each $i \in N$ we have p_i a profit that is collected when the node i is visited. $[O_i, E_i]$ is a time window defining the feasible arrival time at the node and s_i the amount of service time expended for visiting the node i.

This are the end-points of the computed path, where $p_1 = p_n = 0$, $s_1 = s_n = 0$, $a_1 = a_n = 0$ and $b_1 = b_n = T$, with T_{max} represent maximum feasible arrival time at final node, that is $T_{max} = max_{i \in N \setminus 1,n} b_i + s_i + t_{in}$.

The problem a set K of m non-overlapping elementary paths shall be calculated (expect the origin), each path $k \in K$ is an ordered sequence of nodes that start from node 1 and end at node n. Given a solution, v_i indicate the arrival time at node i.

We introduce A a set of directed arcs, such as $\forall i, jE, (i, j), (j, i) \in A$ with $t_{ij} = t_{ji}$. If we now define a binary variable x_{ij}^k and an integer variable that containing the time customer i is visited in path k, $z_i^k = 1$ if the location i is not in path k and it corresponds to v_i otherwise. The problem can be formulated as a Mixed-Integer Programming as follows. First we define the following sets, parameters and variables.

Sets:

N a set of locations
C a set of categories
K a set of routes
I a set of POIs
M a large constant
B a tourist budget
I^c a set of POIs belonging to category c

Parameters :

O_i The opening time on time window
E_i The ending time on time window
t_{max} The maximum time for each visit
t_{ij} The travel time from location i to j
s_i The service duration at location i
p_i The positive profit that is collected when the location i is visited
c_i The cost of the service at location i
Min_c The minimum number of visited POIs of category c
Max_c The maximum number of visited POIs of category c

Variables:

$$x_{ij}^k = \begin{cases} 1 & \text{if arc (i, j)} \in A \text{ is in path } k \\ 0 & \text{otherwise} \end{cases}$$

$$y_i^k = \begin{cases} 1 & \text{if POI } i \text{ is included in path } k \\ 0 & \text{otherwise} \end{cases}$$

$$z_i^k = \begin{cases} 1 & \text{if the location } i \text{ is not in path } k \\ v_i & \text{otherwise} \end{cases}$$

zki= 1if i does not mean anything and it corresponds to otherwise.

The proposed formulation aim to maximize the total collected score S, which enables to maximize the satisfaction of the tourist (maximize the total collected profit):

$$\sum_{k \in K} \sum_{(i,j) \in N} p_i x_{ij}^k \quad (1)$$

itsubject to

$$\sum_{k \in K} \sum_{j \in N} x_{ij}^k \leq 1 \qquad \forall i \in N (2)$$

$$\sum_{j \in N} x_{1j}^k = 1 \qquad \forall k \in K (3)$$

$$\sum_{i \in N} x_{ih}^k - \sum_{j \in N} x_{hj}^k = 0 \qquad \forall h \in N \setminus \{1, n\}, \forall k \in K (4)$$

$$\sum_{i \in N} x_{in}^k = 1 \qquad \forall k \in K (5)$$

$$\mathrm{z}_i^k + t_{ij} + s_i - M(1 - x_{ij}^k) \leq z_j^k \qquad \forall k \in K (6)$$

$$\mathrm{a}_i \leq z_j^k \leq b_i \qquad \forall i \in N, \forall k \in K (7)$$

$$\sum_{j \in N} x_{ij}^k \leq y_j^k \qquad \forall i \in N, \forall k \in K (8)$$

$$\sum_{i \in N} c_i x_{ij}^k \leq B \qquad \forall i \in N (9)$$

$$\sum_{k \in K} y_i^k \leq 1 \qquad \forall i \in N (10)$$

$$Min_c \leq \sum_{i \in I^c} y_i^k \leq \mathrm{Max}_c \qquad \forall k \in K, c \in C (11)$$

$$\mathrm{x}_{ij}^k \in \{0, 1\} \qquad \forall i, j \in A \forall k \in K (12)$$

$$\mathrm{z}_i^k \in Z_+ \qquad \forall i \in P \forall k \in K (13)$$

$$\mathrm{y}_i^k \in \{0, 1\} \qquad \forall i \in V, \forall k \in P (14)$$

- Constraints (2) impose that each customer is visited at most once.
- Constraints (3), (4) and (5) impose that paths have a feasible structure.

- Inequalities (6) and (7) deal with time windows restrictions.
- Constraints (8) impose that a POI can be visited only by a route that she was assigned.
- Constraints (9) ensure that the budget is respected during the travel.
- Constraints (10) impose that each POI can be assigned to at most one route.
- Constraints (11) impose that, for each category c , the number of POIs are visited in each route is between Min_c and Max_c.
- Finally, constraints (12), (13) and (14) define the domain of variables. Note that the special case where $m = |P| = 1$ corresponds to the OPTW.

3. A memetic algorithm for the TOPTW

The Memetic algorithm (MA) [3] is a powerful algorithm that combines an evolutionary algorithm and local search techniques. We present a general architecture of our MA for TOPTW in flowchart 1. The procedure starts with population initialization by creating feasible solutions that will improved with a tabu search procedure, the presented algorithm iterates between two phases, a crossover operator and a mutation operator. For each iteration we select a new population at the end. The process stops when we achieve a maximum number of generations is reached. More components derails of the memetic algorithm are presented in the following sections.

3.1. Creating initial population

Creating the initial population P starts by composing a sequence of solutions $S1, S2, ..., Sp$. For each solution from the initial population we predefine starting and ending nodes. Then we pick randomly a node $i \notin S$ from the obligatory node set O and insert it at the best position in the current solution. This process is stopped wen all the obligatory nodes are included in our solution S.
The $3-opt$ operator is used to shorten the path length of solution. It involves deleting 3 edges from one route, then reconnecting this route in all other possible ways, and finally it select the best path among the new ways [5]. Each newly constructed solution is improved with a tabu search procedure before insertion into the population, and is detailed in the next section.

3.2. Tabu search

The tabu search procedure is important component of our memetic algorithm. At each stage of the improvement, the memetic algorithm selects probabilistically between two evaluation functions, $\Phi 1$ and $\Phi 2$ to allow for violation of the time constraint and the tourist budget.

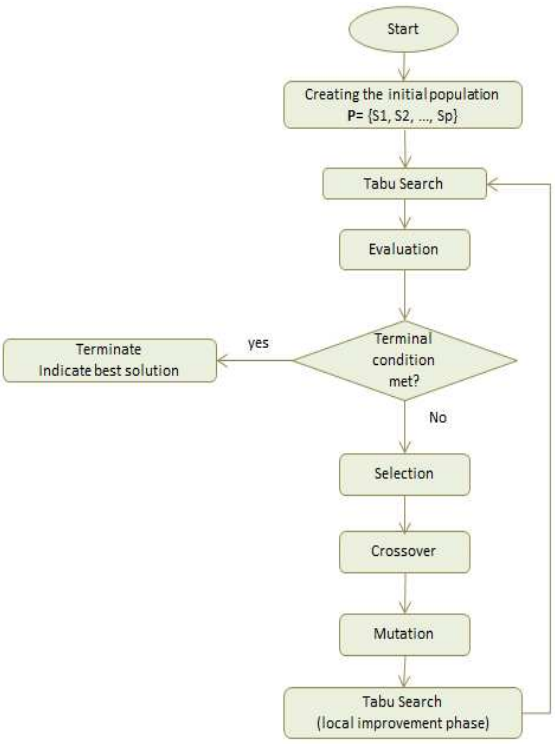


Figure 1. Flowchart of the proposed Memetic algorithm

The main components are detailed thereafter. Starting from a solution S, the iterations of the tabu search consists in determining the best non-tabu solution, with a maximisation of Φ value, By combining the three neighborhoods of non-tabu solution using three move operators (ADD, DROP and SWAP). The three move operators are defined as follows:

- ADD used to insert an optional node at the best position in S that were not included in the current solution
- SWAP used to exchange an optional node which did not exist before in S with an optional existing node in S
- DROP used to delete an optional node from S

A $3-opt$ move operator is only applied once to shorten the length of the Hamiltonian path of the solution [5]. The ϱ parameter reduced by a factor of 2 if feasible solutions have been achieved or increased by a factor of 2 if infeasible solutions have been achieved.
Eventually, the best-found solution S^* is replaced by S if S is feasible and if its higher then S^* in terms of the total collected score.If an improvement for the best solution S^* was not achieved during the iterations, we return S^* as the final output of the tabu search.

3.3. Crossover operator

In our memetic algorithm he crossover operator is applied, at each generation, to create two new offspring

individuals. The combination of two selected parents solutions randomly from the population, and creates two new offspring individuals, that will be added to the population.
The crossover operator based on backbone [28,29], where the subset of nodes shared in the two parents presented by the backbone of s_1 and s_2 ($s_1 \cap s_2$) having respectively offspring o_1 and o_2.
The construction of offspring starts by identifying the subset s' of common nodes in the two parents s_1 and s_2 where $s' = \{\nu : \nu \in s_1, \nu \in s_2\}$.
Then all the nodes from s' ares respectively copied to o_1 and o_2 by keeping the same order in their parents s_1 and s_2. If $t_{(o1)} < t_{max}$, we select a node k from the candidate set $\{\nu : \nu \in I \backslash o_1, \mid O_v \cap o_1 \mid = 0, t(o_1) + e_v \leq t_{max}\}$ then we insert it at the best position in offspring o_1. Repeat the process until no new node can be added while respecting the time constraint violating. And we apply the same process to obtain o_2.

3.4. Mutation operator

If all the nodes from have the same order in both parents, we apply a mutation procedure each newly generated offspring o. The mutation procedure consists in applying random μ Add-Move or Swap-Move to each offspring o, where the parameter μ represent the mutation strength.
When we apply Add-Move the node from the set $\{\nu : \nu \in I \setminus o\}$ is randomly selected and inserted at the best position in offspring o. And when applying Swap-Move, a node from the set $\{\nu : \nu \in I \setminus o\}$ is swapped randomly with randomly selected optional node in o.
Note that the application of Add and Swap moves may result the infeasibility in terms of respecting travel time limit and budget constraints. We eliminate the incompatibility by selecting randomly a node from the candidate set $\{\nu : \nu \in o \setminus M, \mid O_v \cap o \mid > 0\}$ and removing it from o.
Repeat the process until all the nodes in o are compatible. The travel time limit and budget constraints violation is eliminated by randomly deleting an optional node, until a feasible solution is obtained.

3.5. Population updating

As a consequence, the creation of an offspring o and improvement with the tabu search procedure, o replaces the worst solution S_w. The premature convergence represent one potential risk of this classic replacement strategy since o is introduced in the population regardless of its distance to other individuals.
We avoid premature convergence in the proposed memetis algoritme and we ensures a diversified offspring through application of the crossover and mutation procedure described previously.

4. Experimental results

In this section, we describe The computational experiments carried out in our work. The aim of the experiments is to evaluate the accuracy of the proposed Memectic Algorithme to solve the introduced TTDP with clustred POIs.
We shall present all results obtained by using a personal computer with $Intel^{®}\ Core^{TM}$ i5-5200U CPU @ 2.20 GHz 6,00 GO RAM and Windows 10 pro, x64-based processor. The above mathematical model of the TOPTW was coded using CPLEX 12.6. For the proposed Memetic algorithm, JEE has been used.
In the numerical experiment, we tested our algorithms on at set of instances contains 30 instances with 30 POIs. We consider a set 30 POIs clustered into 5 different categories (culture, adventure, family,sea and restaurant). We tested our algo-rithms on two classes of data instances, originally introduced by Montemanni and Gambardella [11] for standard TOPTW.
Several parameters of the TTDP with clustred POIs such as number of categories, the time windows, service duration, minimum and maximum number of visited POIs of each category and the cost of the service at the POI. The process of generating instances considers 5 categories. The data relating to constraints for category was defined based on the specific category.
The generated instances for each category are presented in Table 1.

Category	O_i	E_i	s_i	Min_c-Max_c
Culture	8-9	17	2h	1-2
Adventure	10	20-22	2h-8h	0-1
Family	9-10	22-23	2h-5h	1-3
Sea	8-9	17-19	2h-5h	0-2
Restaurant	12-13	15-16	1h-2h	1-1

TABLE 1. CATEGORIES FEATURES

4.1. Results

To evaluate the performance of the Memetic Algorithme, this is compared with other heuristics of the state of the art for the TOPTW. We use the instances presented in Montemanni and Gambardella [11] and Vansteenwegen et al. [6] for the comparison. The performance of proposed algorithm is compared against the existing approachs: the Ant Colony Optimization(ACO) heuristic of Montemanni and Gambardella [11],the Variable Neighborhood Search (VNS) of Tricoire et al. [9], and the Fuzzy y greedy randomized adaptive search procedure (F-GRASP) of Expósito et al. [22]. Our algorithm is designed for the TTDP with clustred POIs that is a version of the TOPTW with new constraint of tourist budget.

Instance Set	ACO		VNS		F-GRASP		MA	
	rpe	cpu_{avg}	rpe	cpu_{avg}	rpe	cpu_{avg}	rpe	cpu_{avg}
m= 4								
c101	1018	1048,1	1013	78,5	1020	484.8	1025	641,2
c102	1142	1211,3	1139	90,1	1090	166,7	1102	198,5
c108	1112,1	820	1123	70,8	1120	289,5	1124	289,9
c109	1172,1	916	1174	73,8	1150	270,5	1136,9	287,7
r101	608	55,1	610,2	40	605	128,5	599	131,2
r102	825,5	1924,5	828,4	58,4	828	840,2	826,4	977,1
r103	902,5	2622,2	909,8	68,8	854	774,5	875	687
r109	876,7	1374,5	870,5	55,3	832	93,7	824	104,1
r110	900,7	925,3	898,1	70,7	868	181,8	846	174,1
r111	932,4	1596,6	936,6	63,6	912	591,5	898	485,5
r112	947,7	1662,5	964,4	63,3	907	83,7	901,7	345
rc101	805.4	1324,4	777,2	55,3	748	280,9	774,4	297,2
rc102	899,4	2218,8	893,4	67	786,8	698,7	756,4	765,4
rc103	941,8	2005,2	945,4	64,1	936	345,5	918	401,2
rc104	1013,4	2139,3	1033,5	64	1030	373,7	1034,8	397,8
rc105	867,4	1052,3	859	53,6	858	840,4	852,3	911,5
rc16	901,4	2106,7	894,4	49,6	891	443,2	889,2	546,5
rc107	959,2	1763,1	958	53,8	945	1427,9	925,4	1223,4
rc108	1000,4	2222,2	1011,5	59,6	985	333,2	974,8	504,5

TABLE 2. COMPARISON OF RESULTS FOR 4 ROUTES.

5. Conclusion

The tourist trip design problem with clustered POIs while the time and budged constraints is supposed as an NP-hard problem derived from several practical cases. We present a highly effective memetic algorithm (MA) to tackle the problem.

The goal of this work is to mathematically model and the problem and formalize the problem to provide efficient algorithm able to provide a high quality solution. We proposed a mathematical model to optimize problem and solve it. To address these issues, we also propose a Memetic Algorithm improved with a tabu search procedure for the specific features of the problem, with the goal of generating high-quality solutions within reasonable times.

On the basis of the contributions presented in this work, several solid research directions are still open. They include Time depending on the visit, multi-modal transport, hotel selection and developed a support system decision.

References

[1] P. Vansteenwegen and D. Van Oudheusde, The mobile tourist guide: an or opportunity. Oper Res Insight 2007;20(3):21–7.

[2] P. Vansteenwegen . Planning in Tourism and Public Transportation-Attraction Selection by Means of a Personalised Electronic Tourist Guide and Train Transfer Scheduling. PhD thesis, Katholieke Universiteit Leuven 2008.

[3] J. K Hao. Memetic algorithms in discrete optimization. In Handbook of memetic algorithms 2012;73-94.Springer Berlin Heidelberg.

[4] Hu, Q., Lim, A. An iterative three-component heuristic for the team orienteering problem with time windows. Eur. J. Oper. Res. 2014;276-286.

[5] M. Castro.K. Sorensen, P. Vansteenwegen and P. Goos. A memetic algorithm for the travelling salesperson problem with hotel selection. Computers Operations Research 2013;1716-1728.

[6] Vansteenwegen, P., Souffriau, W., Vanden Berghe, G., and Van Oudheusden, D. Iterated local search for the team orienteering problem with time windows. Computers Operations Research 2009;3281-3290.

[7] Liu, L., Xu, J., Liao, S. S., Chen, H. A real-time personalized route recommendation system for self-drive tourists based on vehicle to vehicle communication. Expert Systems with Applications. 2014;3409-3417

[8] Tsai, C.Y., Chung, S.H. A personalized route recommendation service for theme parks using RFID information and tourist behavior. Decision Support Systems 2012;514-527.

[9] Tricoire, F. and Romauch, M. and Doerner, K. F and Hartl, R. F. Heuristics for the multi-period orienteering problem with multiple time windows. Computers & Operations Research. 2010;351-367.

[10] Rodriguez, B., Molina, J., Perez, F., Caballero, R. Interactive design of personalised tourism routes. Tourism Management. 2012;33.92-940.

[11] M.Roberto,, L. M. GAMBARDELLA. An Ant Colony System for team oreanting problem with time windows. Foundation Of Computing And Decision Sciences 2009;287-306.

[12] Souffriau, W., Vansteenwegen, P., Vertommen, J., Berghe, G. V., Oudheusden, D. V. A personalized tourist trip design algorithm for mobile tourist guides. Applied Artificial Intelligence 2008;964-985.

[13] SOUFFIAU, W., MAERVOET, J., VANSTEENWEGEN, P., et al. A mobile tourist decision support system for small footprint devices. International Work-Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg 2009;22(10):1248-1255.

[14] Rodriguez, B., Molina, J., Perez, F., Caballero, R. Interactive design of personalised tourism routes. Tourism Management. 2012;33.92-940.

[15] Golden, B.L., LevyL. , R. Vohra. The orienteering problem, Naval Research Logistics. Naval Research Logistics. 1987;307-318.

[16] Chao, I-M. Chao, Golden, B.L. , and Wasil, E.A. The team orienteering problem. European Journal of Operational Research. 1996;20(3):464-474.

[17] Li, Z., Hu, X. The team orienteering problem with capacity constraint and time window. The 10th International Symposium on Operations Research and its Applications. 2011;157-163.

[18] Kantor M, Rosenwein M. The orienteering problem with time windows. Journal of the Operational Research Society. 1992;629-35.

[19] Righini G, Salani M. Decremental state space relaxation strategies and initialization heuristics for solving the orienteering problem with time windows with dynamic programming. Computers Operations Research. 2009;1191-203.

[20] Souffriau, W. Automated Tourist Decision Support. PhD thesis, Katholieke Universiteit Leuven. 2010.

[21] Labadi, N., Mansini, R., Melechovsky, J., Wolfer Calvo, R. The team orienteering problem with time windows: an lp-based granular variable neighborhood search. Eur. J. Oper. Res. 2012;15-27.

[22] Expósito, A. and Mancini, S. and Brito, J. and Moreno, J. A. A fuzzy GRASP for the tourist trip design with clustered POIs. Expert Systems with Applications. 2019;210–227.

[23] Lin,S.-W.,Yu,V.F. A simulated annealing heuristic for the team orienteering problem with time windows. Eur. J. Oper. Res. 2012;94-107.

An Evidential Spammer Detection based on the Suspicious Behaviors' Indicators

Malika BEN KHALIFA
Université de Tunis,
Institut Supérieur de gestion, LARODEC
Tunis, Tunisia
malikabenkhalifa2@gmail.com

Zied ELOUEDI
Université de Tunis,
Institut Supérieur de gestion, LARODEC
Tunis, Tunisia
zied.elouedi@gmx.fr

Eric LEFEVRE
Univ. Artois, EA 3926,
LGI2A,
62400 Béthune, France
eric.lefevre@univ-artois.fr

***Abstract*—The e-reputation is the key factor for the success of different companies and organizations. It is mainly influenced by the online reviews that have an important impact on the company's development. In fact, they affect the buying decision of the customer. Due to this attraction, the spammers post deceptive reviews to deliberately mislead the potential customers. Thus, the spammer detection becomes crucial to control the fake reviews, to protect the e-commerce from the fraudsters' activities and to ensure an equitable online competition. In this way, we propose a novel method based on the K-nearest neighbor algorithm within the belief function theory to handle the uncertainty involved by the suspicious behaviors' indicators. Our method relies on several spammers indicators used as features to perform the distinguishing between innocent and spammer reviewers. To evaluate our method performance and robustness, we test our approach on two large real-world labeled datasets extracted from yelp.com.**

***Index Terms*—Spammer detection, Online reviews, Fake reviews, Uncertainty, Classification, E-commerce.**

I. Introduction

Nowadays, internet gives the opportunity to people everywhere in the world to express and share their opinions and attitudes regarding products or services. These opinions called online reviews become one of the most important source of information thanks to their availability and visibility. They are increasingly used by both consumers and organizations. Positive reviews usually attract new customers and bring financial gain. However, negative ones damage the e-reputation of different business which lead to a loss. Reviewing has changed the face of marketing in this new area. Due to their important impact, companies invest money to overqualify their product to gain insights into readers preferences. For that, they rely on spammers to usually post deceptive reviews; positive ones to attract new customers and negative ones to damage the competitors' e-reputation. These fraudulent activities are extremely harmful for both companies and readers. Hence, detecting and analyzing the opinion spam becomes pivotal to save the e-commerce and to ensure trustworthiness and equitable competition between different products and services. Therefore, different researchers have given a considerable attention to this challenging problem. In fact, several researches [7], [13], [14], [22], [25] have been devoted to develop performing method capable of spotting fake reviews and stopping these misleading actions. These approaches can be classified into three global categories; spam review detection based on the reviews contents and linguistic features, group spammer detection based on the relational indicators and spammer detection.

Since spammers are the chief responsible of the appearance of deceptive reviews, spotting them is surly one of the most essential task in this field. Several approaches addressed this problem [14] and succeed to achieve significant results. The spammer detection techniques can be divided into two global categories; graph based method and behaviors indicators based methods.

One of the first studies that relies on the graph representation to detect fake reviews was proposed in [31]. This method attempted a spot of fake reviewers and reviews of online stores. This approach is based on a graph model composed by three types of nodes which are reviewers, reviews and stores. The spamming clues are composed through the interconnections and the relationships between nodes. The detection of these clues is based on the trustiness of reviewers, the honesty of reviews and the reliability of stores. Thanks to these three measures the method generates a ranking list of spam reviews and reviewers. This method was tested on real dataset extracted from resellerratings.com and labeled by human experts judged. However, the accuracy of this method is limited to 49%. Similar study was proposed in [11] based also on the review graph model. This method generates a suspicion score for each node in the review graph and updates these scores based on the graph connectivity using an iterative algorithm. This method was performing using a dataset labeled through human judgment. Moreover, the third graph related approach was introduced by [1] as an unsupervised framework. This method relies on a bipartite network composed by reviewers and products. The review can be positive or negative according to the rating. The method assumes that the spammers usually write positive reviews for a bad products and negative ones for good quality products. The authors use an iterative propagation algorithm as well as the correlations between nodes and assign a score to each vertex and update it using the loopy belief propagation (LBP). This method offers a list of scores to rank reviewers and products in order to get k clusters. Results were compared to two iterative classifiers, where they have shown performance.

The aspect of the behaviors indicators was introduced by [18] to detect spammers. This method measures the spamming behaviors and accord a score to rank reviewers regarding the rating they give. It is essentially based on the assumption that fake reviewers target specific products and that their reviews rating deviates from the average rating associated to these products. Authors assume that this method achieved significant results. Another method proposed in [24] is based also on the rating behavior of the each reviewer. It focuses on the gap between the majority of the given rating and each reviewer's rating. This method uses the binomial regression to identify spammers. One of the most preferment studies was proposed by [12], which is essentially based on various spammers behavioral patterns. Since the spammers and the genuine reviewers display distinct behaviors, the proposed method models each reviewer's spamicity while observing his actions. It was formulated as an unsupervised clustering problem in a Bayesian framework. The proposed technique was tested on data from Amazon and proves its effectiveness. Moreover, authors in [12] proposed a method to detect the brust pattern in reviews given to some specific products or services. This approach generates five new spammer behavior indicators to enhance the review spammer detection. The authors used the Markov random fields to model the reviewers in brust and a hidden node to model the reviewer spamicity. Then, they rely on the loopy belief propagation framework to spot spammers. This method achieves 83.7% of precision thanks to the spammers behaviors indicators. Since then, behavioral indicators have become an important basis for spammer detection task. These indicators are used in several recent researches [17]. Nevertheless, we believe that the information or the reviewers' history can be imprecise or uncertain. Also, the deceptive behavior of users might be due to some coincidence which make the spammer detection issue full of uncertainty. For these reasons, ignoring such uncertainty may deeply affect the quality of the detection. To manage these concerns, we propose a novel method aims to classify reviewers into spammer and genuine ones based on K-nearest neighbors' algorithm within the Belief function theory to deal with the uncertainty involved by the spammer behaviors indicators which are considered as features. It is known as the richest theory in dealing with all the levels of imperfections from total ignorance to full certainty. In addition, it allows us to manage different pieces of evidence, not only to combine them but also to make decision while facing imprecision and imperfections. This theory prove its robustness in this field through our previous methods which achieve significant results [3]–[6]. Furthermore, the use of the Evidential K-NN has been based on its robustness in the real world classification problems under uncertainty. We seek to involve imprecision in the spammers behaviors indicators which are considered as the fundamental interest in our approach since they are used as features for the Evidential K-NN. In such way, our method distinguishes between spammers and innocents reviewers while offering an uncertain output which is the spamcity degree related to each user.

This paper is structured as follows: In the first section, we present the basic concepts of the belief function theory and the evidential K-nearest neighbors, then we elucidate the proposed method in section 2. Section 3 is consacred for the experimental results and we finish with a conclusion and some future work.

II. Belief function theory

In section, we elucidate the fundamentals of the belief function theory as well as the Evidential K-nearest neighbors classifier.

A. Basics

The belief function theory, called also the Dempster Shafer theory, is one of the powerful theories that handles uncertainty in different tasks. It was introduced by Shafer [26] as a model to manage beliefs.

1) Basic concepts: In this theory, a given problem is represented by a finite and exhaustive set of different events called the frame of discernment Ω. 2^Ω is the power set of Ω that includes all possible hypotheses and it is defined by: $2^\Omega = \{A : A \subseteq \Omega\}$.

A basic belief assignment (bba) or (a belief mass) represents the degree of belief given to an element A. It is defined as a function $m^\Omega$from 2^Ω to $[0, 1]$ such that:

$$\sum_{A \subseteq \Omega} m^\Omega(A) = 1. \tag{1}$$

A focal element A is a set of hypotheses with positive mass value $m^\Omega(A) > 0$.

Several types of bba's have been proposed [29] in order to model special situations of uncertainty. Here, we present some special cases of bba's:

- The certain bba represents the state of total certainty and it is defined as follows: $m^\Omega(\{\omega_i\}) = 1$ and $\omega_i \in \Omega$.
- The categorical bba has a unique focal element A different from the frame of discernment defined by: $m^\Omega(A) = 1$, $\forall A \subset \Omega$ and $m^\Omega(B) = 0$, $\forall B \subseteq \Omega\ B \neq A$.
- Simple support function: In this case, the bba focal elements are $\{A, \Omega\}$. A simple support function is defined as the following equation:

$$m^\Omega(X) = \begin{cases} w & \text{if } X = \Omega \\ 1 - w & \text{if } X = A \text{ for some } A \subset \Omega \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

  where A is the focus and $w \in$ [0,1].

2) Belief function: The belief function, denoted bel, includes all the basic belief masses given to the subsets of A. It quantifies the total belief committed to an event A by assigning to every subset A of Ω the sum of belief masses committed to every subset of A.

bel is represented as follows:

$$bel(A) = \sum_{\varnothing \neq B \subseteq \Omega} m^\Omega(B) \tag{3}$$

$$bel(\emptyset) = 0 \tag{4}$$

3) Plausibility function: The plausibility function, denoted pl, calculates the maximum amount of belief that could be provided to a subset A of the frame of discernment Ω. Otherwise, it is equal to the sum of the bbm's relative to subsets B compatible with A.

$$pl(A) = \sum_{A \cap B \neq \varnothing} m^{\Omega}(B) \tag{5}$$

4) Combination Rules: Various numbers of combination rules have been proposed in the framework of belief functions to aggregate a set of bba's provided by pieces of evidence from different experts. Let m_1^{Ω} and m_2^{Ω} two bba's modeling two distinct sources of information defined on the same frame of discernment Ω. In what follows, we elucidate the combination rules related to our approach.

1) $Conjunctive\ rule$: It was settled in [30], denoted by $\textcircled{\cap}$ and defined as:

$$m_1^{\Omega} \textcircled{\cap} m_2^{\Omega}(A) = \sum_{B \cap C = A} m_1^{\Omega}(B) m_2^{\Omega}(C) \tag{6}$$

2) $Dempster$**'s** $rule\ of\ combination$: This combination rule is a normalized version of the conjunctive rule [8]. It is denoted by $\oplus$ and defined as:

$$m_1^{\Omega} \oplus m_2^{\Omega}(A) = \begin{cases} \frac{m_1^{\Omega} \textcircled{\cap} m_2^{\Omega}(A)}{1 - m_1^{\Omega} \textcircled{\cap} m_2^{\Omega}(\emptyset)} & \text{if} \quad A \neq \emptyset, \forall A \subseteq \Omega, \\ 0 & otherwise. \end{cases} \tag{7}$$

5) Decision process: The belief function framework provides numerous solutions to make decision. Within the Transferable Belief Model TBM [30], the decision process is performed at the pignistic level where $bba's$ are transformed into the pignistic probabilities denoted by $BetP$ and defined as:

$$BetP(B) = \sum_{A \subseteq \Omega} \frac{|A \cap B|}{|A|} \frac{m^{\Omega}(A)}{(1 - m^{\Omega}(\emptyset))} \quad \forall\ B \in \Omega \tag{8}$$

B. Evidential K-Nearest neighbors

The Evidential K-Nearest Neighbors (EKNN) [9] is one of the best known classification methods based in the belief function framework. It performs the classification over the basic crisp KNN method thanks to its ability to offer a credal classification of the different objects. This credal partition provides a richer information content of the classifier's output.

Notations

- $\Omega = \{C_1, C_2, ..., C_N\}$: The frame of discernment containing the N possible classes of the problem.
- $X_i = \{X_1, X_2, ..., X_m\}$: The object X_i belonging to the set of m distinct instances in the problem.
- A new instance X to be classified.
- $N_K(X)$: The set of the K-Nearest Neighbors of X.

EKNN method

The main objective of the EKNN is to classify a new object X based on the information given by the training set. A new instance X to be classified must be allocated to one class of the $N_K(X)$ founded on the selected neighbors. Nevertheless, the knowledge that a neighbor X belongs to class C_q may be deemed d as a piece of evidence that raises the belief that the object X to be classified belongs to the class C_q. For this reason, the EKNN technique deals with this fact and treats each neighbor as a piece of evidence that support some hypotheses about the class of the pattern X to be classified. .In fact, the more the distance between X and X_i is reduces, the more the evidence is strong. This evidence can be illustrated by a simple support function with a bba such that:

$$m_{X,X_i}(\{C_q\}) = \alpha_0 \exp^{-(\gamma_q^2 d(X,X_i)^2)} \tag{9}$$

$$m_{X,X_i}(\Omega) = 1 - \alpha_0 \exp^{-(\gamma_q^2 d(X,X_i)^2)} \tag{10}$$

Where;

- α_0 is a constant that has been fixed in 0.95.
- $d(X, X_i)$ represents the Euclidean distance between the instance to be classified and the other instances in the training set.
- γ_q assigned to each class C_q has been defined as a positive parameter. It represents the inverse of the mean distance between all the training instances belonging to the class C_q.

After the generation of the different $bba's$ by the K-nearest neighbors, they can be combined through the Dempster combination rule as follows:

$$m_X = m_{X,X_1} \oplus ... \oplus m_{X,X_K} \tag{11}$$

where $\{1, ..., K\}$ is the set including the indexes of the K-Nearest Neighbors.

III. Proposed Method

The idea behind our method is to take into account the uncertain aspect in order to improve detecting the spammer reviewers. For that, we propose a novel approach based on different spammers indicators and we rely on the Evidential K-nearest neighbors which is famous classifier under the belief function framework. In the remainder of this section we will elucidate the different steps of our proposed approach; in the first step we model and calculate the spammers' indicators through the reviewers' behaviors. In the second step, we present the initialization phase. Moreover, the learning phase is detailed in the third step. Finally, we distinguish between the spammers and the innocent reviewers through the classification phase in which we also offer an uncertain input to report the spamicity degree of each reviewer. Figure I illustrates our method steps.

A. Step1: Pre-processing phase

As mentioned before, the spammers indicators become one of the most powerful tool in the spammers detection field used in several researches. In this part, we propose to control the reviewers behaviors if they are linked with the spamming activities and thus can be used as features to learn the Evidential KNN classifier in order to distinguish between

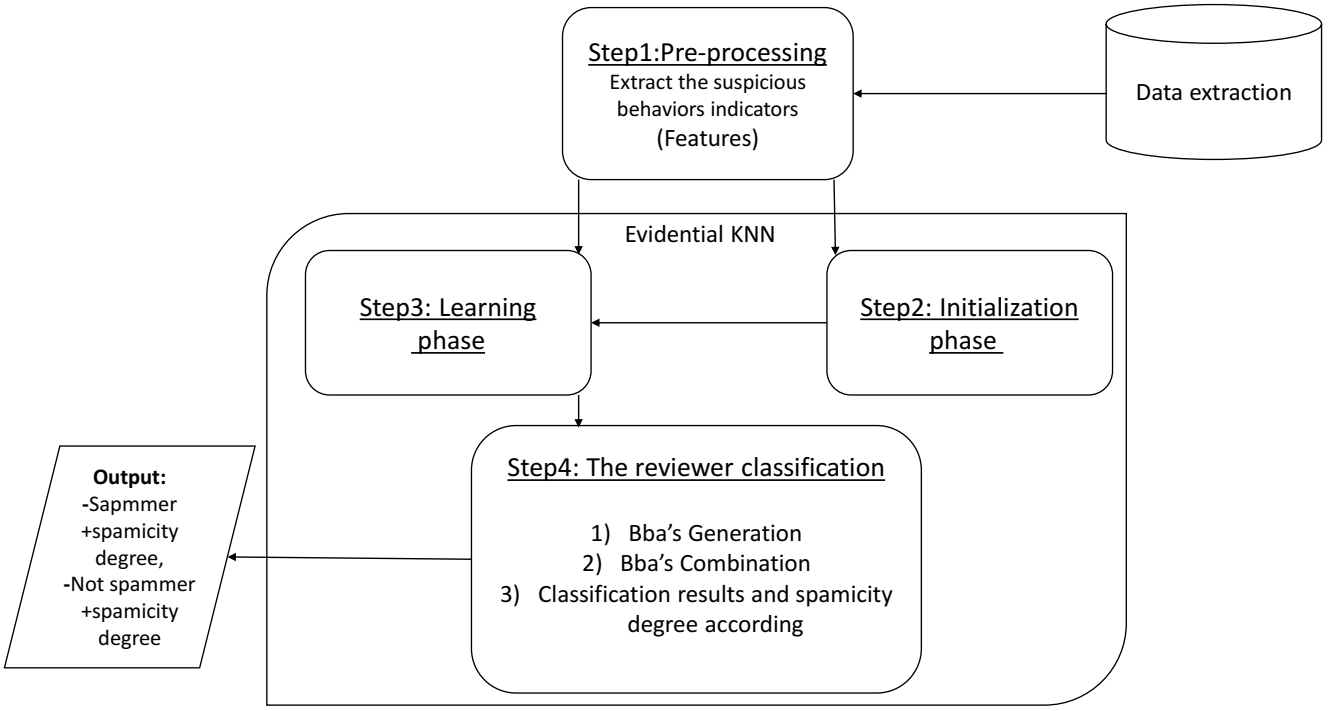


Fig. 1. Our method illustration

the two classes spammer and innocent reviewers. We select the significant features used in the previous work [20]. Here, we detail them in two lists; in the first list we elucidate the author features and the second one presents the review features. To make the equations more comprehensible we present the different notations in the table I.

Reviewers features: The values of these features are into the interval [0,1]. The more the value is close to 1 the higher the spamicity degree is indicated.

1) Content similarity (CS): Generally, spammers choose to copy reviews from other similar products because, for them, creating a new review is considered as an action that required time. That's why, we assume that it is very useful to detect the reviews' content similarity (using cosine similarity) of the same reviewer. From this perspective and in order to pick up the most unpleasing behavior of spammers, we use the maximum similarity.

$$f_{CS}(R_i) = max_{R_i(r_j),R_i(r_k)\in R_i(Tr)} cosine(r_j, r_k) \quad (12)$$

Where $R_i(r_j)$ and $R_i(r_k)$ are the reviews written by the reviewer R_i, and $R_i(Tr)$ represents all the reviews written by the reviewer R.

2) Maximum Number of Reviews (MNR): Creating reviews and posting them successively in one day display an indication of a deviant behavior. This indicator calculates the maximum number of reviews per day for a reviewer normalized by the maximum value for our full data.

$$f_{MNR}(R_i) = \frac{MaxRev(R_i)}{Max_{R_i\in R_i(Tr)} MaxRev(R_i)} \quad (13)$$

3) Reviewing Burstiness (BST) : Although authentic reviewers publish their reviews from their accounts occasionally, the opinion spammers represent a non-old-time membership in the site. To this point, it makes us able to take advantage of the account's activity in order to capture the spamming behavior. The activity window, which is the dissimilarity between the first and last dates of the review creation, is used as a definition of the reviewing burstiness. Consequently, if the time-frame of a posted reviews was reasonable, it could mention a typical activity. Nevertheless, posting reviews in a short and nearby burst (τ= 28 days, estimated in [20]), shows an emergence of a spam behavior.

$$f_{BST}(R_i) = \begin{cases} 0 & L(R_i(r)) - F(R_i(r)) > \tau \\ \frac{L(R_i(r)) - F(R_i(r))}{\tau} & Otherwise \end{cases} \quad (14)$$

Where $L(R_i(r))$ represents the last posting date of the review r given by the reviewer R_i and $F(R_i(r))$ is first posting date of the review.

4) Ratio of First Reviews (RFR): To take advantage of the reviews, people lean on the first posted reviews. For this reason, spammers tend to create them at an early stage in order to affect the elementary sales. Therefore, spammers believe that managing the first reviews of each product could empower them to govern the people's sentiments. For every single author, we calculate the ratio between the first reviews and the total reviews. We mean by the first reviews those

TABLE I
LIST OF NOTATION

R_i	A reviewer
r	A review
p	A product
Tr	Total number of reviews
$R_i(r)$	Review written by the reviewer R_i
$R_i(Tr)$	Total number of reviews written by the reviewer R_i
$Tr(p)$	Total number of reviews on product or service p
$r(p)$	Review on product p
$R_i(r(p))$	Review given by the reviewer R_i to the same product p
$R_i(Tr(p))$	Total number of reviews given by the reviewer R_i to the same product p
$R_i(Tr_*(p))$	Total number of rating reviews given by the reviewer R_i to the same product p
$L(R_i(r))$	Last posting date of the review written by the reviewer R_i
$F(R_i))$	First posting date of the review written by the reviewer R_i
$A(p)$	The date of the product launch
I_i	Spamming indicator
S_{mean}	The mean score of a given product
S	The reviewing score of the reviews given to one product p by the same reviewer R_i.
$\Omega = \{S, \bar{S}\}$	The frame of discernment including the spammer and not spammer class

posted by the author as the first to evaluate the product.

$$f_{RFR}(R) = \frac{|R_i(r_f) \in R_i(Tr)|}{R_i(Tr)} \tag{15}$$

Where $R_i(r_f)$ represents the first review of the reviewer R_i.
Review features: These features have a binary values. If the feature value is equal to 1, then it indicates the sapmming. If not, it represents the non-spamming.

5) Duplicate/Near Duplicate Reviews (DUP): As far as they want to enhance the ratings, spammers frequently publish multiple reviews. They tend to use a duplicate/near-duplicate kind of preceding reviews about the same product. We could spotlight this activity by calculating the duplicate reviews on the same product. The calculation proceeding is as following:

$$f_{DUP}R_i(r)) = \begin{cases} 1 & r \in R_i(Tr(p)) = cosine(R_i(r), r) > \beta_1 \\ 0 & otherwise \end{cases} \tag{16}$$

For a review r each author R_i on a product p acquires as value 1 if it is in analogy (using cosine similarity based on some threshold, $\beta_1 = 0.7$) with another review is estimated in [20].

6) Extreme Rating (EXT): In favor of bumping or boosting a product, spammers often review it while using extreme ratings (1* or 5*). We have a rating scale composed by 5 stars (*).

$$f_{EXT}(R) = \begin{cases} 1 & R_i(Tr_*(p)) \in \{1, 5\} \\ 0 & R_i(Tr_*(p)) \in \{2, 3, 4\} \end{cases} \tag{17}$$

Where $R_i(Tr_*(p))$ represents all the reviews (ratings) given by the reviewer R_i to the same product p.

7) Rating Deviation: Spammers aim to promote or demote some target products or services to this point they generate reviews or rating values according the situation. In order to deviate the overall rating of a product, they have to contradict the given opinion by posting deceptive ratings strongly deviating the overall mean.
If the rating deviation of a review exceeds some threshold $\beta 2 = 0.63$ estimated in [20], this features achieves the value of 1. The maximum deviation is normalized to 4 on a 5-star scale.

$$f_{Dev}(R) = \begin{cases} 1 & \frac{|S - S_{mean}|}{4} > \beta_2 \\ 0 & otherwise \end{cases} \tag{18}$$

Where S_{mean} represents the mean score of a given product and S represents the reviewing score of the reviews given to one product p by the same reviewer R_i.

8) Early Time Frame (ETF): Since the first review is considered as a meaningful tool to hit the sentiment of people on a product, spammers set to review at an early level in order to press the spam behavior. The feature below is proposed as a way to detect the spamming characteristic:

$$ETF(r, p) = \begin{cases} 0 & L(R_i, p) - A(p) > \delta \\ 1 - \frac{L(R_i,p) - A(p)}{\delta} & otherwise \end{cases} \tag{19}$$

$$f_{ETF}(r) = \begin{cases} 1 & ETF(R_i, R_i(r(p))) > \beta_3 \\ 0 & otherwise \end{cases} \tag{20}$$

Where $L(R_i, p)$ represents the last review posting date by the reviewer R_i on the product p and $A(p)$ is the date of the product launch. The degree of earliness of an author R_i who had reviewed a product p is captured by $ETF(R_i, R_i(r(p)))$ the threshold symbolizing earliness is about $\delta = 7$ months

(estimated in [20]). According the presented definition, we cannot consider the last review as an early one if it has been posted beyond 7 months since the product's launch. On the other hand, the display of a review following the launch of the product allows this feature to reach the value of 1. $\beta_3 = 0.69$ is considered as the threshold mentioning spamming and is estimated in [20].

9) Rating Abuse (RA): To bring up the wrongly use generated from the multiple ratings we adopt the feature of Rating Abuse (RA). Obtaining Multiple rating on a unique product is considered as a weird behavior. Despite the fact that this feature is alike to DUP, it does not focus on the content but rather it targets the rating dimension. As definition, the Rating Abuse, the similarity of the donated ratings by an author for a product beyond multiple ratings by the same author blended by the full reviews on this product.

$$RA(R_i, R_i(r(p)) = |R_i(Tr(p))|(1 - \frac{1}{4} max_{r \in R_i(Tr(p))}(r,p) - min_{r \in R_i(Tr(p))}(r,p)) \quad (21)$$

$$f_{RA} = \begin{cases} 1 & RA(R_i, R_i(r(p)) > \beta_4 \\ 0 & otherwise \end{cases} \quad (22)$$

We should calculate the difference between the two extremes (maximum/minimum) on 5-star scale rating to catch the coherence of high/low rating and to determine the similarity of multiple star rating. The maximum difference between ratings attains as normalized constant 4. Lower values are reached by this feature if, in authentic cases, the multiple ratings where in change (as a result of a healthy use). $\beta_4 = 2.01$ is considered as the threshold mentioning spamming and is estimated in [20].

B. Step2: Initialization phase

In order to apply the Evidential K-NN classifier, we should firstly assign values to parameters α_0 et γ_0 to be used in the learning phase. We will start by initializing the parameter α_0 and then computing the second parameter γ_{Ii} while exploiting the reviewer-item matrix. As mentioned in the EKNN procedure [9], the α_0 is initialized to 0.95. The value of the parameter α_0 is assigned only one time while the γ_{Ii} value change each time according to the current items' reviewers. In order ensure the γ_{Ii} computation performance, first of all we must find reviewers having separately exclusive spammers indicators. Based on the selected reviewers, we assign a parameter γ_{Ii} to each indicators I_i corresponding to the reviewer R_i which will be measured as the inverse of the average distance between each pair of reviewers R_i and R_j having the same spammers' indicators values. This calculation is based on the Euclidean distance denoted $d(R_i, R_j)$ such that:

$$d(R_i, R_j) = \sqrt{\sum_{i,j=1}^{n} (I_{(R,i)} - I_{(R,j)})^2} \quad (23)$$

Where $I_{(R,i)}$ and $I_{(R,j)}$ correspond to the value of the spammer indicators of the reviewer R to the indicators i and j.

C. Step3: Learning phase

Once the spammers indicators are calculated and the two parameters α_0 and γ_{Ii} have been assigned, we must select a set of reviewers. Then, we compute for each reviewer R_j in the database, its distance with the target reviewer R_i. Given a target reviewer, we have to spot its K-most similar neighbors, by selecting only the K reviewers having the smallest distances values that is calculated using the Euclidean distance and denoted by $dist(R_i, R_j)$.

D. Step4: Classification phase

In this step, we aim to classify a new reviewer into spammer or innocent reviewer. Let $\Omega = \{S, \bar{S}\}$ where S represents the class of the spammers reviewers and $\bar{S}$ includes the class of the not spammers (genuine) reviewers.

1) The bba's generation: Each reviewer R_I induces a piece of evidence that builds up our belief about the class that he belongs. However, this information does not supply certain knowledge about the class. In the belief function framework, this case is shaped by simple support functions, where only a part of belief is committed to $\omega_i \in \Omega$ and the rest is assigned to Ω. Thus, we obtain the following bba:

$$m_{R_i,R_j}(\{\omega_i\}) = \alpha_{R_i} \quad (24)$$

$$m_{R_i,R_j}(\Omega) = 1 - \alpha_{R_i} \quad (25)$$

Where R_i is the new reviewers and R_j is its similar reviewer that $j = \{1..K\}$, $\alpha_{R_i} = \alpha_0 \exp^{(-\gamma_{Ii} dist(R_i,R_j))}$, α_0 and γ_{Ii} are two parameters assigned in the initialization phase and $dist(R_i, R_j)$ is the distance between the two reviewers R_i and R_j computed in the learning phase.
In our case, each neighbor of the new reviewer has two possible hypotheses. It can be similar to a spammer reviewer in which his the committed belief is allocated to the spammer class S and the rest to the frame of discernment Ω. In the other case, it can be near to an innocent reviewer where the committed belief is given to the not spammer class $\bar{S}$ and the rest of is assigned to Ω. We treat the K-most similar reviewers as independent sources of evidence where each one is modeled by a basic belief assignment. Hence, K different bba's can be generated for each reviewer.

2) The bba's combination: After the generation of the bba's for each reviewer R_i, we describe how to aggregate these bba's in order to get the final belief about the reviewer classification. Under the belief function framework, such bba's can be combined using the Dempster combination rule. Therefore, the obtained bba represent the evidence of the K-nearest Neighbors regarding the class of the reviewer. Hence, this global mass function m is obtained as such:

$$m_{R_i} = m_{R_i,R_1} \oplus m_{R_i,R_2} \oplus .... \oplus m_{R_i,R_K} \quad (26)$$

3) Final classification result and the spamicity degree according: We apply the pignistic probability $BetP$ in order to select the membership of the reviewer R_i to one of the classes of Ω and to accord him a spamicity degree. Then, the classification decision is made either the reviewer is a

TABLE II
DATASETS DESCRIPTION

Datasets	Reviews (filtered %)	Reviewers (Spammer %)	Services (Restaurant or hotel)
YelpZip	608,598 (13.22%)	260,277 (23.91%)	5,044
YelpNYC	359,052 (10.27%)	160,225 (17.79%)	923

TABLE III
COMPARATIVE RESULTS

Evaluation Criteria	Accuracy				Precision				Recall			
Methods	NB	SVM	UCS	Our method	NB	SVM	UCS	Our method	NB	SVM	UCS	Our method
YelpZip	60%	65%	78%	**84%**	57%	66%	76%	**85%**	63%	68%	74%	**86%**
YelpNYC	61%	68%	79%	**85%**	62%	69%	79%	**86%**	61.8%	67.8%	76.7%	**83.6%**

spammer or not. For this, we select the $BetP$ with the grater value. Moreover, we assign to each reviewer even he is not a spammer the spamicity degree which consists on the $BetP$ value of the spammer class.

IV. EXPERIMENTATION AND RESULTS

The evaluation in the fake reviews detection problem was always a challenging issue due to the unavailability of the true real world growth data and variability of the features also the classification methods used by the different related work which can lead to unsafe comparison in this field.

Data description

In order to test our method performance, we use two datasets collected from yelp.com. These datasets represent the more complete, largest, the more diversified and general purpose labeled datasets that are available today for the spam review detection field. They are labeled through the classification based on the yelp filter which has been used in various previous works [3], [4], [13], [21], [25] as ground truth in favor of its efficient detection algorithm based on experts judgment and on various behavioral features. Table II introduces the datasets content where the percentages indicate the filtered fake reviews (not recommended) also the spammers reviewers. The YelpNYC dataset contains reviews of restaurants located in New York City; the Zip dataset is bigger than the YelpNYC datasets, since it includes businesses in various regions of the U.S., such that New York, New Jersey, Vermont, Connecticut and Pennsylvania. The strong points of these datasets are:

- The high number of reviews per user, which facilities to modeling of the behavioral features of each reviewer.
- The miscellaneous kinds of entities reviewed, i.e., hotels and restaurants
- Above all, the datasets hold just fundamental information, such as the content, label, rating, and date of each review, connected to the reviewer who generated them. With regard to considering over-specific information, this allows to generalize the proposed method to different review sites.

Evaluation Criteria

We rely on these three following criteria to evaluate our method: Accuracy, precision and recall and they can be defined as Eqs.27, 28, 29 respectively where TP, TN, FP, FN denote True Positive, True Negative, False Positive and False Negative respectively:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{27}$$

$$Precision = \frac{TP}{(TP + FN)} \tag{28}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{29}$$

Experimental results

As our method relies on the Evidential KNN classifier to classify the reviewer into spammer and genuine ones. We propose to compare our method with the Support Vector Machine (SVM) and the Naive Bayes (NB) used by most of spammer detection method [17], [20], [25]in this field. Moreover, we propose to compare also with our previous proposed Uncertain Classifier to detect Spammers (UCS) in [5]. Table III reports the different results.

Our method achieves the best performance detection according to accuracy, precision and recall over-passing the baseline classifier. We record at best an accuracy improvement over 24% in both yelpZip and yelpNYC data-sets compared to NB and over 19% compared to SVM. Moreover, the improvement records between our two uncertain methods (over 10%) at best, shows the importance of the variety of the features used in our proposed approach.

Our method can be used in several fields by different reviews websites. In fact, these websites must block the detected spammers in order to stop the appearance of the fake reviews. Moreover and thanks to our uncertain output which represent

the spamicity degree for each reviewer, they can control the behavior of the genuine ones with a high spamicity degree to prevent their tendency to turn into spammers.

V. Conclusion

In this work, we tackle the spammer review detection problem and we propose a novel approach that aims to distinguish between the spammer and the innocent reviewers while taking into account the uncertainty in the different suspicious behavioral indicators. Our method shows its performance in detecting the spammers reviewers while according a spamicity degree to each reviewer. Our proposed approach can be useful for different reviews sites in various fields. Moreover, our uncertain input can be used by other methods to model the reliability each reviewer. As future work, we aim to tackle the group spammer aspect in the interest of improving the detection in this field.

References

[1] Akoglu, L., Chandy, R., Faloutsos, C.: Opinion fraud detection in online reviews by network effects. Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM, 13, 2-11 (2013)

[2] Bandakkanavar, RV., Ramesh, M., Geeta, H.: A survey on detection of reviews using sentiment classification of methods. IJRITCC, 2(2):310–314 (2014) Advanced Intelligent System and Informatics (AISI), 395-404 (2018)

[3] Ben Khalifa, M., Elouedi, Z., Lefèvre, E. Multiple criteria fake reviews detection based on spammers' indicators within the belief function theory. The 19th International Conference on Hybrid Intelligent Systems (HIS'2019). Springer International Publishing. (To appear)

[4] Ben Khalifa, M., Elouedi, Z., Lefèvre, E. Fake reviews detection based on both the review and the reviewer features under belief function theory. The 16th international conference Applied Computing (AC'2019), 123-130 (2019)

[5] Ben Khalifa, M., Elouedi, Z., Lefèvre, E. Spammers detection based on reviewers' behaviors under belief function theory. The 32nd International Conference on Industrial, Engineering Other Applications of Applied Intelligent Systems (IEA/AIE'2019). Springer International Publishing, 642-653 (2019)

[6] Ben Khalifa, M., Elouedi, Z., Lefèvre, E. Multiple criteria fake reviews detection using belief function theory. The 18th International Conference on intelligent systems design and applications (ISDA'2018). Springer International Publishing, 315-324 (2018)

[7] Deng, X., Chen, R.: Sentiment analysis based online restaurants fake reviews hype detection. Web Technologies and Applications, 1-10 (2014)

[8] Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. Ann. Math. Stat.38, 325-339 (1967)

[9] Denoeux, T.: A K-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Trans. Syst. Man Cybern. 25(5), 804–813 (1995)

[10] Lefèvre, E., Elouedi, Z.: How to preserve the confict as an alarm in the combination of belief functions? Decis. Support Syst.56, 326-333 (2013)

[11] Fayazbakhsh, S., Sinha, J.: Review spam detection: A network-based approach. Final Project Report: CSE 590 (Data Mining and Networks) (2012)

[12] Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM, 13, 175-184 (2013)

[13] Fontanarava, J., Pasi, G., Viviani, M.: Feature Analysis for Fake Review Detection through Supervised Classification. Proceedings of the International Conference on Data Science and Advanced Analytics, 658-666 (2017).

[14] Heydari, A., Tavakoli, M., Ismail, Z., Salim, N.: Leveraging quality metrics in voting model based thread retrieval. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 10 (1), 117-123 (2016)

[15] Jindal, N., Liu, B.: Opinion spam and analysis. Proceedings of the 2008 International Conference on Web Search and Data Mining. ACM, pp. 219-230 (2008).

[16] Jousselme, A.-L., Grenier, D., Bossé, É.: A new distance between two bodies of evidence. Inf. Fusion 2(2), 91-101 (2001)

[17] Liu, P., Xu, Z., Ai, J., Wang, F.: Identifying Indicators of Fake Reviews Based on Spammers Behavior Features," IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 396–403 (2017)

[18] Lim, P., Nguyen, V., Jindal, N., Liu, B., Lauw, H. : Detecting product review spammers using rating behaviors. Proceedings of the 19th ACM international conference on information and knowledge management, 939-948 (2010)

[19] Ling, X., Rudd, W.: Combining opinions from several experts. Applied Artificial Intelligence an International Journal, 3 (4), 439-452 (1989)

[20] Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M.: Spotting opinion spammers using behavioral footprints. Proceedings of the ACM international conference on knowledge discovery and data mining, 632-640 (2013)

[21] Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: What Yelp Fake Review Filter Might Be Doing. Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM, 409-418 (2013)

[22] Ong, T., Mannino, M., Gregg, D.: Linguistic characteristics of shill reviews. Electronic Commerce Research and Applications, 13 (2), 69-78 (2014)

[23] Pan, L., Zhenning, X., Jun, A., Fei, W.: Identifying indicators of fake reviews based on spammer's behavior features. Proceedings of the IEEE International Conference on Software Quality, Reliability and Security Companion, QRS-C, 396-403 (2017)

[24] Savage, D., Zhang, X., Yu, X., Chou, P., Wang, Q.: Detection of opinion spam based on anomalous rating deviation. Expert Systems with Applications, 42 (22), 8650-8657 (2015)

[25] Rayana, S., Akoglu, L.: Collective opinion spam detection: Bridging review networks and metadata. Proceedings of the 21th International Conference on Knowledge Discovery and Data Mining, ACM SIGKDD, 985-994 (2015)

[26] Shafer, G.: A Mathematical Theory of Evidence, vol. 1. Princeton University Press (1976)

[27] Smets, P.: The combination of evidence in the transferable belief model. IEEE Trans. Pattern Anal. Mach. Intell. 12(5), 447-458 (1990)

[28] Smets, P.: The transferable belief model for expert judgement and reliability problem. Reliability Engineering and system safety, 38, 59-66 (1992)

[29] Smets, P.: The canonical decomposition of a weighted belief. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 1896-1901 (1995)

[30] Smets, P.: The transferable belief model for quantified belief representation. In: Smets, P. (ed.) Quantified Representation of Uncertainty and Imprecision, 267-301. Springer, Dordrecht (1998)

[31] Wang, G., Xie, S., Liu, B., Yu, P. S.: Review graph based online store review spammer detection. Proceedings of 11th international conference on data mining, ICDM, 1242-1247 (2011)

Managing uncertainty during CBR systems vocabulary maintenance using Relational Evidential C-Means

Safa BEN AYED
Université de Tunis, Institut Supérieur de Gestion, LARODEC, Tunis, Tunisia
Univ. Artois, EA 3926, LGI2A, 62400 Béthune, France
Email: safa.ben.ayed@hotmail.fr

Zied ELOUEDI
Université de Tunis, Institut Supérieur de Gestion, LARODEC, Tunis, Tunisia
Email: zied.elouedi@gmx.fr

Eric LEFEVRE
Univ. Artois, EA 3926, LGI2A, 62400 Béthune, France
Email: eric.lefevre@univ-artois.fr

***Abstract*—Due to the incremental learning of Case-Based Reasoning (CBR) systems, there is a colossal need to maintain their knowledge containers which are *(1)* the case base, *(2)* similarity measures, *(3)* adaptation, and *(4)* vocabulary knowledge. Actually, the vocabulary presents the basis of all the other knowledge containers since it is used for their description. Besides, CBR systems store real-world experiences which are full of uncertainty and imprecision. Therefore, we propose, in this paper, a new policy to maintain vocabulary knowledge using one of the most powerful tools for uncertainty management called the belief function theory, as well as the machine learning technique called Relational Evidential C-Means (RECM). We restrict the vocabulary knowledge to be the set of features describing cases, and we aim to eliminate noisy and redundant attributes by taking into account the correlation between them.**

1. Introduction

Case-Based Reasoning (CBR) is a methodology based on solving new problems using past experiences. Actually, each new target problem triggers an entire cycle; In a nutshell, the CBR system *(a) retrieves* from the Case Base (CB) the most similar case(s). Then, it *(b) reuses* that case so as to propose an adapted solution. However, this solution can be rejected, so it should be *(c) revised*. Finally, the confirmed solution with its corresponding problem will be retained in the CB to serve, as a new case, for future problems resolution [1] (see Figure 1).

Since CBR systems proved over the years a widespread interest in several domains, and since they are designed to work over long time frames, their Knowledge Containers (KC) [2] are now the subject of a considerable maintenance targets. There are four KCs in CBR systems: (1) *CB*, (2) *Vocabulary*, (3) *Similarity measures* and (4) *Adaptation rules*. Their maintenance turns around the field of Case-Based Reasoner Maintenance (CBRM) [3]. In fact, research are maily focused on Case Base Maintenance (CBM) [4], [5], [6], [7], [8] as well as models to estimate CBs competence in problem solving [9], [10]. However, there is no deep exploration on maintaining the vocabulary container although it presents the basis of all the other knowledge containers [2]. Actually, it needs maintenance for different reasons like context modification and domain overfitting. The most prevalent structure of vocabulary knowledge within CBR is the attribute-value representation [2]. Otherwise, other structures may be used such as predicates, set of words, and related constructed. In this work, we present the vocabulary knowledge as the set of features describing cases. Hence, it is relevant to maintain them by keeping only those that improve the competence of CBR systems in problem-solving. Generally, there are two kinds of knowledge that researchers aim to eliminate during maintenance: noisiness and redundancy. Noisy attributes present feature's knowledge that distorts problems solving operations. Redundant attributes have no added-value in offering accurate solutions, and their elimination will improve the performance of the

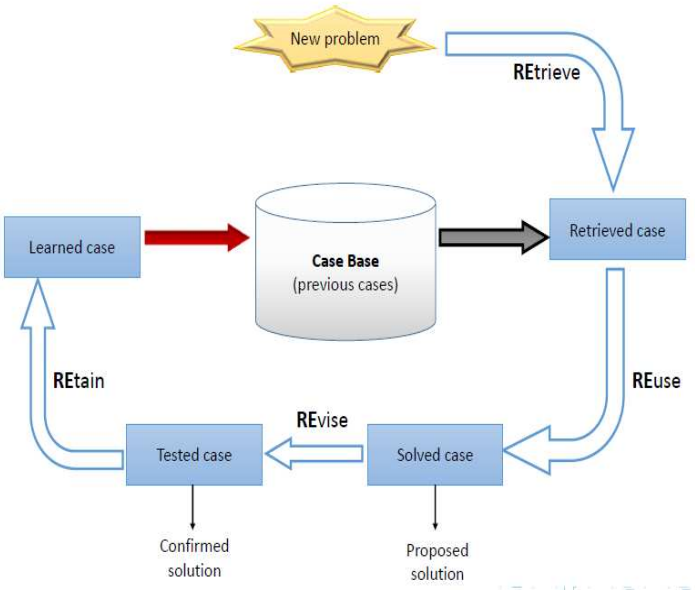


Figure 1. The CBR cycle

CBR system.

To select relevant features, *Attribute Clustering* concept [12], [23], [24] has proved high adequacy in CBR context since it allows to maintain the relation between features, as well as providing a flexibility in term of features substitution. Most of vocabulary maintenance policies are hard i.e they are not able to manage uncertainty within knowledge. Undoubtedly, there is a crucial need to handle imperfection within CBR systems since embedded knowledge refer to real-world experiences, which are generally uncertain and imprecise. Belief function theory or Evidence theory [25], [26] presents one of the most powerful tools for that matter. Some evidential clustering techniques such as k-evclus and EVCLUS [27], [28], have been used in [12] and [29] for features learning. However, other techniques [30] within this theory have been proposed and showed their effectiveness. Therefore, we aim, in this paper, to propose a new policy for vocabulary knowledge maintenance based on assessing the relation between features through correlation measurement and using the Relational Evidential C-Means (RECM) [30] as a machine learning technique. We also aim to show results offered by our policy and compare them to policies using other methods, as well as to the original non-maintained CBR.

The rest of the paper is organized as follows. The next Section defines the vocabulary container, overviews applicable notions for vocabulary maintenance, and present some related works. Section 3 presents tools and fundamental concepts of the belief function theory. During Section 4, we describe in details our proposal for maintaining vocabulary maintenance. Results will be discussed in Section 5. Finally, we end by concluding within Section 6.

2. Maintaining knowledge within CBR systems

As already mentioned, CBR systems contain different varieties of knowledge that are stored in four knowledge containers, which are Case Base, Similarity measures, Adaptation knowledge, and Vocabulary. These knowledge needs maintenance along the time on account of appearance of noisiness and redundancy as well because context change (see Figure 2).

Before moving and focusing on the literature of vocabulary knowledge and its maintenance to introduce our contribution, for the current work, let briefly present the other three knowledge containers within CBR systems.

Case Base Maintenance. Firstly, to maintain the CB knowledge container which defines the set of stored experiences, we find, in the literature, several policies that are based on different strategies to maintain cases such as:

1) The *selection based strategy* which contains CBM policies that revise CB's content through the selection of only representative cases that are able to cover the remaining set of cases problems. For instance, the Condensed Nearest Neighbor (CNN) [13] presents the baseline of data reduction methods.

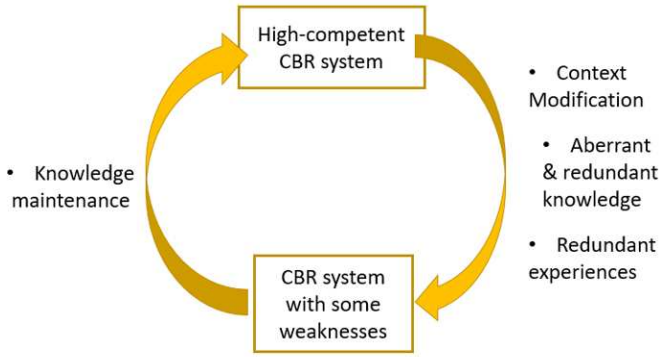


Figure 2. CBR knowledge maintenance

2) The *optimization based strategy* which contains policies that apply maintenance operations according to some evaluation criteria such that:
   - *The performance* which is quantified by the time spent to solve a target problem,
   - *The competence* which represents the range of problems that the CB can successfully solves [14]

   Among that policies, Iterative Case Filtering algorithm (ICF) [15] is based on the competence criterion to make decision about cases deletion.
3) The *partitioning based strategy* which is characterized by giving the ability of treating the original CB in form of small ones, which makes it easier and more effective. Generally, the different subsets of cases are obtained using the clustering as an unsupervised machine learning technique. Actually, cases clustering is widely used within the CBM field due to its approved utility in detecting cases to be maintained. For instance, Evidential Clustering and case Types Detection for case base maintenance method (ECTD) [5] performs cases evidential clustering technique and, then, selects only cases that their deletion affects the whole CB quality. Some variants of ECTD policies have also been proposed [6], [8].

Similarity Maintenance. Secondly, Similarity measures are mainly used in CBR during the *Retrieve* step[1]. In the literature, there are several works focusing on the maintenance of similarity, and especially on learning features weighting in similarity measures [16], [17].

1. In the current research, we use the Euclidean Distance as a retrieve similartity measure inside the nearest neighbor algorithm (it is one of the most used algorithms in CBR systems).

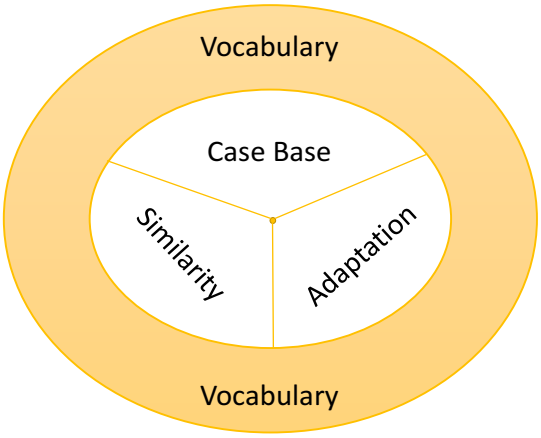


Figure 3. Illustrating the four knowledge containers within CBR systems

According to authors in [18], there are three main maintenance operations if we limit ourselves to weighted linear measures: modify a weight, modify a local measure, and stretch out a measure to a new attribute.

Adaptation Maintenance. Thirdly, within this adaptation maintenance area, research has to be so careful in adaptation changes. In fact, a low-quality maintenance may extremely influence the performance of the CB, and consequently the whole CBR system.

Handling adaptation rules is a very intensive knowledge. Consequently, we generally appeal domain-experts interventions for this matter. Nevertheless, there are different works in the context of adaptation learning and maintenance. For instance, an automated mechanism has been proposed in [19] to learn some adaptation rules from the different observations in the CB: If descriptions of two cases vary in just few number of features, the core of the adaptation rule is then formed from the differences in those features. Other works are also carried out in this area such that [20], [21]. Let us mention that, according to [22], there are four types of adaptations that can be applied inside CBR systems:

1) *Null Adaptation*[2]: It consists on returning the similar case solution without any modification,
2) *Transformational Adaptation*: It applies a modification in some parts of the solution description,
3) *Generative Adaptation*: It generates a solution from scratch.
4) *Compositional Adaptation*: It combines the three previous types to obtain the reused solution.

Vocabulary Maintenance. According to authors in containers [11], and as shown in Figure 3, the vocabulary is presented as the basis of all the other knowledge. During this Section, we define the vocabulary and the reason of maintenance, along with mentioning some related work and applicable concepts for maintenance.

2. This corresponds to the type of adaptation considered by the current research work.

2.1. Vocabulary container definition

To start, the vocabulary may be defined as the response to the question "Which element of the data structures are used to present fundamental notions?" [2]. In fact, the definition of this knowledge is highly depended on knowledge source nature. For object-oriented organization, for instance, attribute-value structure is generally used to define the vocabulary. However, if it consists on more complex types of data like text, sensor data or image, then the vocabulary may be defined differently using predicates, functions, set of words, or related constructs.

In the current work, we define the vocabulary of CBR systems by the set of features[3] describing cases.

2.2. The need of vocabulary knowledge maintenance

Obviously, there is a need to maintain all knowledge containers within CBR systems since they are designed to work for a long period of time. For instance, let assume that we possess a CBR system with high competence and performance. After a period of time, we may encounter noisy and redundant knowledge as well as some context changes. Hence, the system will know some weaknesses and degradation of competence and/or performance.

In this paper, we are focusing only on maintaining the set of features. In fact, every experience, in our life, can be described with unlimited number of features, where only some of them are considered as representative and able to direct the best solution. Generally, two kinds of features should be removed; noisy attributes which overwrite the smooth learning, and redundant ones which slow CBR system's operations.

2.3. Applied concepts for vocabulary maintenance

To maintain CBR vocabulary knowledge, different concepts within the machine learning field have been applied. Among the most used ones for this matter, we cite feature selection and attribute clustering.

Firstly, since we work with structured CBR systems and aim to retain only relevant attributes, the field of *Feature Selection* (FS) is suitable for such problem. FS is a NP-Hard problem that aims to select only relevant attribute, for some data, that do not contribute to the predictive model's accuracy. Hence, we note some FS methods that have been combined with CBR vocabulary maintenance context [31], [32], and others that select attributes through allocating weights according to features significance [33].

Secondly, the *Attribute Clustering* (AC) is a suitable way that could be used to maintain vocabulary within CBR context. In fact, AC allows to preserve relation between features which offers a high flexibility to CBR framework since each attribute could be substituted by another one belonging to the same cluster. Consequently, AC has been

3. *Feature* and *Attribute* terms are used exchangeably within this paper.

used in some works [34], [35] as a feature selection method for vocabulary maintenance [12].

The concept of AC is similar to objects clustering where similar objects should belong to the same cluster, and inversely. However, similarity between attributes is described in term of relation between them, such as correlation or dependency, which is generally depending on the task objective.

To manage uncertainty within knowledge, from the complete ignorance to the total certainty, we choose to use tools from the belief function theory [25], [26], as presented in the following Section.

3. Belief function theory

Since our work, elaborated in this paper, aims to manage knowledge imperfection which is naturally embedded within real-world experiences stored in CBR systems, some tools and techniques from the belief function theory are used, in this work, to allow a high-quality vocabulary maintenance. During this Section, the fondamental concepts of the belief function theory, which is named also Demspter-Shafer or Evidence theory [25], [26], are presented. Besides, the concept of credal partition within the evidential clustering is defined to model the doubt about features assignment to clusters.

3.1. Fundamental Concepts

To model and quantify the evidence under the belief function framework, let consider ω a variable that refers to C elementary events for some problem presented by $\Omega = \{\omega_1, \omega_2, ..., \omega_C\}$. We call the Ω set by the *frame of discernment* and the set of all the 2^C possible subsets of the predefined events taking values in Ω by the *power set*. The power set is therefore defined as follows:

$$2^\Omega = \{\emptyset, \{\omega_1\}, ..., \{\omega_C\}, \{\omega_1, \omega_2\}, ..., \{\omega_1, \omega_C\}, ..., \Omega\} \tag{1}$$

The main key point of this theory is called the basic belief assignment (bba) which refers to the partial knowledge regarding the real value of ω. It is defined by the following function:

$$\begin{aligned} m : 2^\Omega &\rightarrow [0, 1] \\ B &\mapsto m(B) \end{aligned} \tag{2}$$

where m satisfies the following constraint:

$$\sum_{B \subseteq \Omega} m(B) = 1 \tag{3}$$

We call an element $B \subseteq \Omega$ as focal if $m(B) > 0$. The set of all the focal elements is called *Body of Evidence (BoE)*. If each element in BoE is a singleton, then m is called a *Bayesian bba*. Otherwise, if BoE contains only the frame of discernment Ω as a focal element, then m is named a *vacuous belief function*. However, if it contains only one focal element of Ω which is singleton, then m is a *Certain mass function*.

The basic belief mass $m(B)$, which presents the degree of belief assigned to the hypothesis "$\omega \in B$", can be attached to a subset of variables regardless any additive assumption. A bba corresponds to the open world assumption when we allow the assignment of evidence to the empty set partition. That means that Ω is incomplete and the actual value may be taken outside the frame of discernment. This interpretation is meaningful especially in clustering when we aim to distuinguish noisy knowledge [30], [37]. Contrariwise, if the bba is normalized ($m(\emptyset) = 0$), then it corresponds to the closed-world assumption [26]. If we are interested to move on from unnormalized ($m(\emptyset) \neq 0$) to normalized ($m(\emptyset) = 0$) bba, then the *Dempster Normalization* can be applied, which consists on dispersing the belief's degrees assigned to the empty set over all the other focal sets such that:

$$m_*(B) = \begin{cases} \dfrac{m(B)}{1 - m(\emptyset)} & if\ B \neq \emptyset \\ 0 & Otherwise \end{cases} \tag{4}$$

Some useful functions are generally computed through the bba for some reason. For instance, for a given bba m, the corresponding belief (bel), plausibility (pl) and commonality (q) functions are from 2^Ω to $[0, 1]$ and defined by [26] as follows:

$$bel(B) = \sum_{\emptyset \neq D \subseteq B} m(D) \qquad \forall B \subseteq \Omega \tag{5}$$

,

$$pl(B) = \sum_{B \cap D \neq \emptyset} m(D) \qquad \forall B \subseteq \Omega \tag{6}$$

and

$$q(B) = \sum_{B \subseteq D} m(D) \qquad \forall B\Omega \tag{7}$$

In other words, the belief function $bel(B)$ represents the total belief that one commits to B without committed to $\overline{B}$, the plausibility function $pl(B)$ quantifies the maximum amount of belief that could be given to a subset B and the commonality function $q(B)$ represents the total mass that is free to move to every element of B.

The source of evidence providing bbas are often not fully reliable. Hence, a *discounting operation* [26] is necessary to update the bba according to the degree of trust assigned to the source. In fact, the discount rate, denoted $\alpha \in [0, 1]$, refers to the amount of belief that expert's information is reliable. The updated bba ${}^\alpha m$ is defined as follows:

$${}^\alpha m(B) = \begin{cases} (1 - \alpha)\ m(B) & for\ B \neq \Omega \\ \alpha + (1 - \alpha)\ m(\Omega) & for\ B = \Omega \end{cases} \tag{8}$$

Let consider two mass functions m_1 and m_2 defined in the same frame of discernment Ω. Hence, the degree of conflict between them may be computed with various ways. One of the most cummon methods is defined, in [26], as follows: $\kappa = \sum_{A \cap B = \emptyset} m_1(A)\ m_2(B)$. Authors, in [27],

demonstrated that if two mass functions m_1 and m_2 quantify evidence regarding two different questions that take values in the same possible answers Ω, then the plausibility that both questions have the same answer is equal to $1 - \kappa$.

At the end, we generally need to make decision after handling beliefs' degrees. Hence, we present one of the most known tools for decision making called the pignistic probability transformation ($BetP$). It consists on choosing the hypothesis, regarding a normalized bba m, having the highest value. $BetP$ is therefore defined as follows:

$$BetP(\omega) = \sum_{\omega \in B} \frac{m(B)}{|B|} \qquad \forall \omega \in \Omega \tag{9}$$

To be able to apply the $BetP$ transformation within the open world assumption ($m(\emptyset) \neq 0$), a prefatory step of normalization (Equation 4) has to be performed.

3.2. Credal Partition

Handling imperfection using the evidence theory within the clustering problem has known a widespread interest in several work. The clustering is a machine learning technique that aims to organize data according to the similarity between instances. The more the objects are similar, the more they intend to share the same group.

The key point of the evidential clustering problem is known by the *Credal Partition*, which is close to the fuzzy partition concept but more general. The credal partition is created by assigning degrees of belief not only to singletons of the frame of discernment, but also to all possible subsets of Ω. Within the evidential clustering, the frame of discernment refers to the set of C possible clusters. The uncertainty towards the membership of an object i to a partition of clusters B is presented by a mass function denoted $m_i(B)$. Its value supports the hypothesis saying "The actual cluster of instance i belongs to the partition B". That bba quantifies the uncertainty regarding the membership of only one object. When there is n instances, the credal partition will be the set of $n - tuple$ bbas $(m_1, m_2, ..., m_n)$.

For a credal partition $M = (m_1, m_2, ..., m_n)$, the two following particular cases are of interest [30]

- If every m_i is a certain bba, then M presents a crisp partition of the frame of discernment Ω, which corresponds to a complete knowledge situation.
- If every m_i is a Bayesian bba, then M presents a fuzzy partition of the frame of discernment of Ω.

4. Vocabulary maintaining process using Relational Evidential C-Means (RECM)

To remove noisy and redundant attributes from vocabulary knowledge, our proposed policy follows three principle steps as presented in Figure 4, and referred by three arrows. It is mainly based on the RECM machine learning technique that is able to handle relations between features and manage uncertainty within data.

4.1. Step 1: Generating relational matrix through studying the correlation between features

Let define the relation between features inversely proportional to the correlation between them. In fact, we aim to express this relation in term of dissimilarity, where the correlation may refer to the similarity between features. Actually, the more two given attributes are correlated, the more they offer the same information and considered as similar. To measure the linear association between two attributes A_i and A_j, we use the Pearson's correlation coefficient [36], which is defined such that:

$$r_{A_iA_j} = \frac{\sum_{l=1}^{n} (a_{il} - \overline{a_i})\ (a_{jl} - \overline{a_j})}{\sqrt{\sum_{l=1}^{n} (a_{il} - \overline{a_i})^2}\ \sqrt{\sum_{l=1}^{n} (a_{jl} - \overline{a_j})^2}} \tag{10}$$

where a_{il} and a_{jl} refer respectively to values of attributes A_i and A_j regarding case l, where $\overline{a_i}$ and $\overline{a_j}$ present their mean values.

Knowing that $r_{A_iA_j}$ is bounded in $[-1, 1]$, three main situations arise to define the relational matrix.

- If $r_{A_iA_j} \simeq 1$, then there is a high correlation (positive) $\Rightarrow$ similar provided information $\Rightarrow$ high similarity.
- If $r_{A_iA_j} \simeq -1$, then there is a high correlation (negative) $\Rightarrow$ similar provided information $\Rightarrow$ high similarity.
- If $r_{A_iA_j} \simeq 0$, then there is no correlation $\Rightarrow$ different provided information $\Rightarrow$ high dissimilarity.

As a consequence, the Relational matrix R is defined as:

$$R = (1 - |r_{A_iA_j}|) \qquad i, j = 1..p \tag{11}$$

where p presents the total number of features.

4.2. Step 2: Learning on attributes using Relational Evidential C-Means (RECM)

During this step, we aim to consider features as objects and cluster them according to their relational matrix previously generated in order to build clusters, where each one contains similar and correlated attributes. Since knowledge are never exact, we express clusters' membership through degrees of belief. To do, we use a powerful technique for this matter called the Relational Evidential C-Means (RECM) [30].

RECM [30] is a relational version of the basic Evidential C-Means (ECM) [37], where both are based on an alternate minimization scheme. However, ECM handles vectorial attribute data and RECM handle dissimilarity data.

Let start by presenting ECM and consider n objects described in p feature space, v_i is the center that represents a given cluster c_i, and $\overline{v_j}$ is the barycenter that represents the partition of clusters C_j with $C_j \subseteq \Omega$. To derive the credal

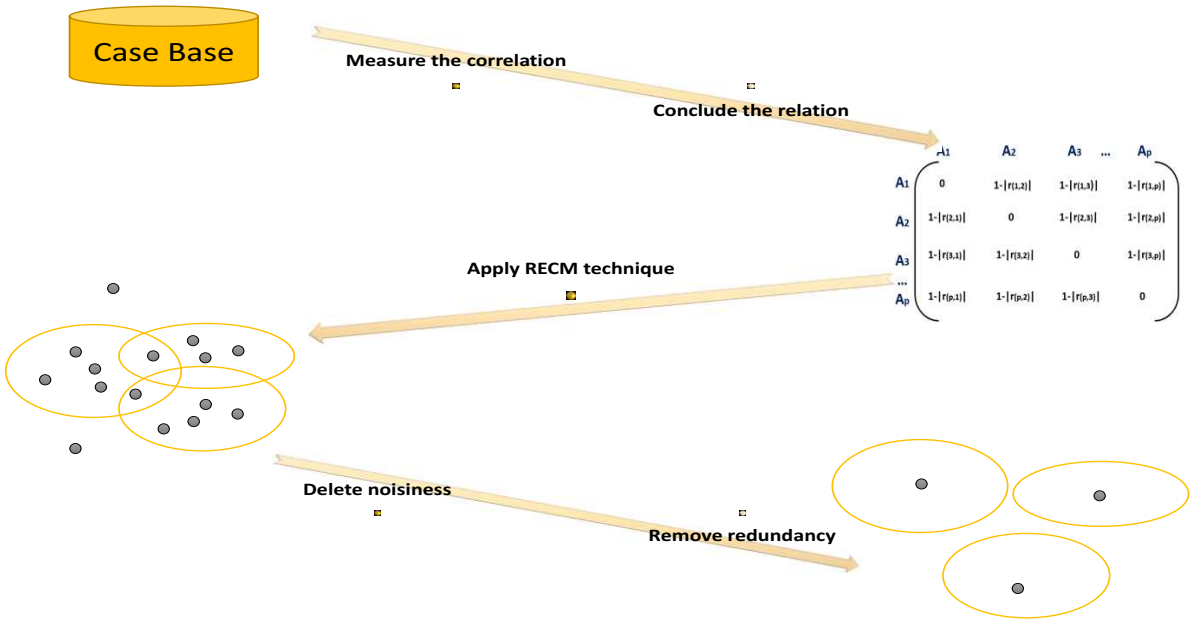


Figure 4. Steps of Evidential vocabulary maintenance based on RECM technique

partition $M = (m_1, ..., m_n)$ and the resulted clusters, which are presented by their centers V, ECM algorithm proceeds an alternate optimization scheme to optimize the following cost function:

$$J_{ECM}(M,V) = \sum_{i=1}^{n} \sum_{j/C_j \neq \emptyset, C_j \subseteq \Omega} |C_j|^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta} \tag{12}$$

subject to

$$\sum_{j/C_j \subseteq \Omega, C_j \neq \emptyset} m_{ij} + m_{i\emptyset} = 1 \qquad \forall i = 1...n \tag{13}$$

where m_{ij} denotes $m_i(C_j)$ and $\emptyset$ refers to noise cluster [30], [37] to be at a fixed distance δ from every instance. The exponent α aims at controlling the penalization degree for partitions having high cardinality, while β and δ treat noisiness.

Firstly, as presented in [37], V is considered as fix and M is updated as follows:

$$m_{ij} = \frac{|C_j|^{-\alpha/(\beta-1)} d_{ij}^{-2/(\beta-1)}}{\sum_{C_k \neq \emptyset} |C_k|^{-\alpha/(\beta-1)} d_{ik}^{-2/(\beta-1)} + \delta^{-2/(\beta-1)}} \tag{14}$$

and

$$m_{i\emptyset} = 1 - \sum_{C_j \neq \emptyset} m_{ij} \qquad \forall i = 1...n \tag{15}$$

Secondly, M is considered fix, and we obtain an unconstrained optimization problem. The resulted V is defined as the solution of the following equation:

$$HV = UX \tag{16}$$

where X represents objects data, and H and U are defined, respectively, as follows:

$$H_{lk} = \sum_{i} \sum_{C_j \neq \emptyset, C_j \supseteq \{\omega_l, \omega_k\}} |A_j|^{\alpha-2} m_{ij}^{\beta} \qquad k,l = 1...C \tag{17}$$

and

$$U_{li} = \sum_{C_j \ni \omega_l} |A_j|^{\alpha-1} m_{ij}^{\beta} \qquad l = 1...C \quad i = 1..n \tag{18}$$

with x_{iq} is the value of feature q regarding object o_i,

For RECM, which is applied in our context, we handle dissimilarity data between features. Hence the notation n with ECM will refer here to the number of features p. Let, in this step, denote the already generated relational matrix R by $\Delta = (\delta_{ii'})$ and consider the square matrix $W = (w_{ii'})$ referring to dot products of features, which is defined as follows:

$$W = -\frac{1}{2} J \Delta J \tag{19}$$

where $J = \frac{1}{n} ee^t - I$ with $e = (1, ..., 1)^t \in \mathbb{R}$ and I is $n \times n$ identity matrix.

Last but not least, let consider the matrix $Q = (q_{kk'})$ (respectively the matrix $Z = (z_{kk'})$) the matrix of dot products of centers (respectively the matrix of doc products between centers and instances). As demonstrated in [30], Z and Q are obtained by solving systems of linear equations toward the following equations respectively:

$$HZ_{.i} = UW_{.i}, \tag{20}$$

and

$$HQ_{.i} = UZ_{i.} \quad (21)$$

Ultimately, let present, in Algorithm 1, the different steps of RECM technique, as presented in [30], by making use of the previous defined Equations.

Algorithm 1 RECM algorithm [30]

Require: - Dissimilarity data;
- The number of clusters C;
- The parameters α, β, δ, and ϵ;

Ensure: - The credal partition M;
- Clusters centers V;

1: **BEGIN**
2: Randomly generate the initial credal partition $M^{(0)}$;
3: Initialization: $k \longleftarrow 0$;
4: Calculate W using Equation 19;
5: **Repeat**
6: $k \longleftarrow k + 1$
7: Calculate $H^{(k)}$ and $U^{(k)}$ from $M^{(k-1)}$ using Equations 17 and 18;
8: **For** i=1..n
9: Calculate the i^{th} column of $Z^{(k)}$ using Equation 20;
10: **End For**
11: **For** i=1..C
12: Calculate the i^{th} column of $Q^{(k)}$ using Equation 21;
13: **End For**
14: Calculate distances d_{ij} from $R^{(k)}$ and $Q^{(k)}$;
15: Update $M^{(k)}$ using Equations 14 and 15;
16: **Until** $||M^{(k)} - M^{(k-1)}|| < \epsilon$
17: **END**

4.3. Step 3: Removing noisy and redundant attributes

To reach our objective, we aim, in this step, to keep only representative features by eliminating noisiness and redundancy. We call noisy features those that have a degree of belief to the empty set higher that all the other degrees. They, hence should be removed since they may gravely reduce the competence of the CBR system to solve problems.

Eliminating redundant features consists on selecting only one representative attribute from every cluster and removing the others. We choose to select the nearest feature to the centre of cluster to which it belongs. The decision of attributes membership to clusters has been made using the pignistic probability transformation as defined in Equation 9.

4.4. Illustrative example

Let consider knowledge regarding vocabulary container is defined by the set attributes A_i with i goes from 1 to 4. Let assume now that the frame of discernment Ω contains two clusters (ω_1 and ω_2) and the applied RECM algorithm offers a credal partion $M = [m_1; m_2; m_3; m_4]$, where its values are presented in Table 1.

TABLE 1. EXAMPLE OF CREDAL PARTITION VALUES

M	$\emptyset$	$\{\omega_1\}$	$\{\omega_2\}$	Ω
m_1	0.05	0.75	0.15	0.05
m_2	0.65	0.1	0.1	0.15
m_3	0.1	0.05	0.8	0.05
m_4	0.2	0.1	0.5	0.2

TABLE 2. PIGNISTIC PROBABILITY TRANSFORMATION VALUES

	ω_1	ω_2
$\boldsymbol{BetP_1}$	**0.8158**	**0.1842**
~~$BetP_2$~~	~~0.5~~	~~0.5~~
$\boldsymbol{BetP_3}$	**0.0833**	**0.9167**
$BetP_4$	0.25	0.75

From values given in Table 1, we remark that $m_2(\emptyset) > m_2(\{\omega_1\}) + m_2(\{\omega_2\}) + m_2(\{\omega_1, \omega_2\})$. Hence, we flag the feature A_2 as noisy according to specifications given by our policy in step 3. By this way, we refresh the vocabulary knowledge by eliminating A_2. Thereafter, we make a decision regarding the membership of attributes to clusters using Equation 9 and we obtain results as shown in Table 2.

We note that A_1 belongs to ω_1, and A_3 and A_4 belong to ω_2. Ultimately, we only retain A_1 as an attribute prototype of ω_1 and A_3 as representative of ω_2 to build the new maintained vocabulary knowledge.

5. Experimentation and results

The main purpose of our experimentation part is to compare results with those offered when we use other evidential learning tools for vocabulary maintenance, as well as to show the effectiveness degree of our proposal that we call REVM for Evidential Vocabulary Maintenance based on RECM [30] technique.

5.1. Data and experimental settings

The different policies and tools, presented in this paper, have been implemented using the *R software*. The original and the maintained CBR systems are tested using six real-world datasets from UCI repository for machine learning[4]. The description of these datasets are presented in Table 3.

During the implementation of some policies, some parameters should be set. The number of clusters (or features p) used by RECM, for every CB, is fixed similarly to those in [12] with the EVM policy. The initial credal partition is randomly set, and the parameter α regarding the cardinality, within function J_{RECM}, is set to 1 which means that we do not penalize clusters' partitions with high cardinality.

4. https://archive.ics.uci.edu

TABLE 3. CASE BASES DESCRIPTION

	Case base	Ref	# instances	# attributes
1	Ionosphere	IO	351	34
2	Glass	GL	214	10
3	WDBC	BC	569	31
4	German	GR	1000	20
5	Heart	HR	270	13
6	Yeast	YS	1484	8

5.2. Testing strategy and evaluation criteria

During the evaluation, we make use of the most applied classification algorithm within the CBR context: k-Nearest Neighbor (k-NN)[5].

Two evaluation criteria have been used, in our work, to assess our proposal efficiency. First, the accuracy criterion consists on measuring the competence of the system in term of the Percentage of Correct Classifications (PCC), and defined such as:

$$PCC(\%) = \frac{\#\ Correct\ classifications}{\#\ Total\ classifications} \times 100 \quad (22)$$

To obtain final estimation of accuracy, we make use of the 10-fold cross validation technique as shown in Figure 5.

Second, we use the retrieval time (RT) criterion which refers to the time spent to retrieve and classify problems by the CBR system.

5.3. Results and discussion

As shown in Tables 4 and 5, we compare results offered by our proposal that we call REVM, denoting *RECM based Evidential Vocabulary Maintenance policy*, to those given by the non-maintained CBR system (Original-CBR), by a feature selection method called ReliefF [33] (CBR-ReliefF), and by the EVM policy [29] which uses EVCLUS [27] [28] for learning.

In term of accuracy, we remark, from Table 4, that RECM method offers competitive results, especially comparing to the non maintained CBR system. In fact, we remark that it was able to improve their initial competence rates provided by Original-CBR with all the tested datasets. For instance, it improves the accuracy of "Yest" dataset from 55.32 % (Original-CBR) to 98.98 % (RECM).
Comparing to the feature selection method called reliefF (ReliefF-CBR), we note, also, that RECM offers close results. Ultimately, we note that very competitive results are provided with both EVM and our current proposed method RECM. Actually, we note that they are slightly in favor with the Evidential Vocabulary Maintenance method (EVM) which uses the k-EVCLUS technique for learning. Although there is no high difference between these results (they offer almost same accuracy values for four datasets among six), we conclude that k-EVCLUS was more suitable. It may be explained by its high flexibility to handle different types of similarity or relational data, since our strategy considers them to be in term of correlations between features. However, we may tolerate this fact since our current proposal was able to retain or even largely improve the initial competence values.

5. We chose to set k to 5 in order to avoid the sensibility to noisiness.

TABLE 4. ACCURACY [PCC(%)]

Case bases		Accuracy [PCC(%)]			
		Original-CBR	ReliefF-CBR	EVM	REVM
1	IO	85.48 %	84.88 %	88.33 %	86.98 %
2	GL	97.64 %	98.11 %	98.59 %	98.12 %
3	BC	60.16 %	96.33 %	96.46 %	96.18 %
4	GR	64.6 %	73.4 %	73.25 %	73.26 %
5	HR	57.5 %	62.45 %	62.98 %	60.91 %
6	YS	55.32 %	99.05 %	99.05 %	98.98 %

TABLE 5. RETRIEVAL TIME [T(S)]

Case bases		Retrieval time [T(s)]			
		Original-CBR	ReliefF-CBR	EVM	REVM
1	IO	1.942	1.188	0.912	0.902
2	GL	0.967	0.882	0.762	0.604
3	BC	1.710	1.112	1.013	0.902
4	GR	1.812	1.213	1.211	1.222
5	HR	2.103	1.091	1.028	0.789
6	YS	0.954	0.722	0.724	0.776

In term of retrieval time, we note that the decreasing of the number of instances as well of features may conduct to decrease the research time, which is described in Table 5 in seconds. We remark that results offered after performing vocabulary maintenance are lower than those offered with the original non maintained CBR systems (Original-CBR). With "Ionosphere", "Breast Cancer", and "Heart" datasets, for instance, we note faster cases retrieval with values go, respectively, from 1.942s to 0.902s, from 1.710s to 0.902s, and from 2.103s to 0.789s. However, we also note very close results with the other ReliefF and EVM maintaining methods, which is logic since they offer data described with close number of features. In fact we would like to mention that, it could be affected by tasks executed in the

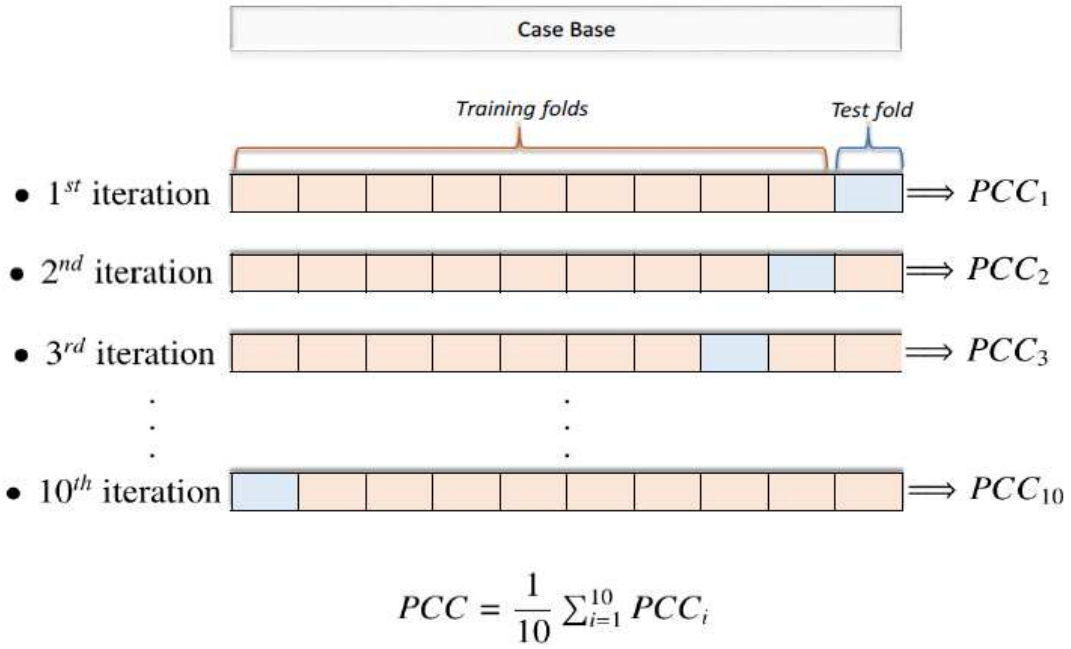


Figure 5. The principle of 10-fold cross validation

environment of development during experimentation.

6. Conclusion

In this paper, we used the Relational Evidential C-Means (RECM) as a machine learning technique at the aim to maintain vocabulary knowledge within CBR systems by keeping only representative features. The idea consists on studying, first of all, the relation between features through measuring the correlation between them and create the relational matrix. This matrix will then be used to learn on features with managing uncertainty using RECM. Ultimately, the generated credal partition has been analyzed and studied to retain only representative and relevant features to describe case knowledge. During the experimentation, we tested our proposal on six real datasets from different domains provided in UCI repository for machine learning. Offered results showed some improvement of problem-solving competence comparing to those offered by the non maintained CBR systems. Good results are also noted comparing to the feature selection method called ReliefF. However, we remarked very competitive results with the vocabulary maintenance policy, called Evidential Vocabulary Maintenance (ECM), that uses the EVCLUS technique for learning.

References

[1] Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. *In Artificial Intelligence Communications*, pp. 39-52 (1994).

[2] Richter, M. M., Michael, M.: Knowledge containers. *Readings in Case-Based Reasoning* (2003).

[3] Wilson, D. C., Leake, D. B.: Maintaining Case-Based Reasoners: Dimensions and Directions. *Computational Intelligence*, pp. 196-213 (2001).

[4] Smiti, A., Elouedi, Z.: SCBM: soft case base maintenance method based on competence model. *Journal of Computational Science, 25*, pp. 221-227 (2017).

[5] Ben Ayed, S., Elouedi, Z., Lefevre, E.: ECTD: evidential clustering and case types detection for case base maintenance. *In IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1462-1469. IEEE (2017).

[6] Ben Ayed, S., Elouedi, Z., Lefevre, E.: DETD: Dynamic Policy for Case Base Maintenance Based on EK-NNclus Algorithm and Case Types Detection. *In International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 370-382 (2018).

[7] Smiti, A., Elouedi, Z.: WCOID-DG: An approach for case base maintenance based on Weighting, Clustering, Outliers, Internal Detection and Dbsan-Gmeans. *Journal of computer and system sciences, 80(1)*, pp. 27-38 (2014).

[8] Ben Ayed, S., Elouedi, Z., Lefevre, E.: Exploiting Domain-Experts Knowledge Within an Evidential Process for Case Base Maintenance. *In International Conference on Belief Functions*, pp. 22-30 (2018).

[9] Ben Ayed, S., Elouedi, Z., Lefevre, E.: CEC-Model: A New Competence Model for CBR Systems Based on the Belief Function Theory. *International Conference on Case-Based Reasoning*, pp. 28-44 (2018).

[10] Smyth, B., McKenna, E.: Modelling the competence of case-bases. *In: Smyth, B., Cunningham, P. (eds.) EWCBR. LNCS, vol. 1488*, pp. 208-220 (1998).

[11] Roth-Berghofer, T. R. and Cassens, J.: Mapping goals and kinds of explanations to the knowledge containers of case-based reasoning systems. *In International Conference on Case-Based Reasoning*, pp. 451-464 (2005).

[12] Ben Ayed, S., Elouedi, Z., Lefevre, E.: Maintaining case knowledge vocabulary using a new Evidential Attribute Clustering method. *In 13th International FLINS Conference on Data Science and Knowledge Engineering for Sensing Decision Support*, pp. 347-354 (2018).

[13] Hart, P.: The condensed nearest neighbor rule. *IEEE transactions on information theory, 14(3)*, pp. 515-516 (1968).

[14] Smyth, B., McKenna, E.: Competence models and the maintenance problem. *Computational Intelligence, 17(2)*, pp. 235-249 (2001).

[15] Brighton, H., Mellish, C.: On the consistency of information filters for lazy learning algorithms. *In Proceedings of european conference on principles of data mining and knowledge discovery*, pp. 283-288 (1999).

[16] Stahl, A., Gabel, T.: Using evolution programs to learn local similarity measures. *In International Conference on Case-Based Reasoning,* pp. 537-551 (2003).

[17] Stahl, A.: Learning similarity measures: A formal view based on a generalized CBR model. *In International Conference on Case-Based Reasoning*, pp. 507-521 (2005).

[18] Richter, M. M., Weber, R. O.: Case-Based Reasoning. *Springer-Verlag Berlin Heidelberg* (2013).

[19] Leake, D. B., Kinley, A., Wilson, D.: Acquiring case adaptation knowledge: A hybrid approach. *In Proceedings of the Thirteenth National Conference on Artificial Intelligence, AAAI Press*, pp. 684-689 (1996).

[20] Jarmulak, J., Craw, S., Rowe, R.: Using case-base data to learn adaptation knowledge for design. *In International joint conference on artificial intelligence, Vol. 17, No. 1*, pp. 1011-1020 (2001).

[21] Craw, S., Jarmulak, J., Rowe, R.: Learning and applying case-based adaptation knowledge. *In International Conference on Case-Based Reasoning*, pp. 131-145 (2001).

[22] Wilke, W., Bergmann, R.: Techniques and knowledge used for adaptation during case-based problem solving. *In Proceedings of the international conference on industrial, engineering and other applications of applied intelligent systems*, pp. 497-506 (1998).

[23] Hong, T. P., Liou, Y. L.: Attribute clustering in high dimensional feature spaces. *In International Conference on Machine Learning and Cybernetics, Vol. 4*, pp. 2286-2289 (2007).

[24] Maji, P.: Fuzzyrough supervised attribute clustering algorithm and classification of microarray data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(1)*, pp.222-233 (2011).

[25] Dempster, A. P.: Upper and lower probabilities induced by a multivalued mapping. *In The annals of mathematical statistics*, pp. 325-339 (1967).

[26] Shafer, G.: A mathematical theory of evidence. *Princeton University Press*, Princeton (1976).

[27] Denœux, T., Masson, M. H.: EVCLUS: Evidential clustering of proximity data. *IEEE Trans. on Systems, Man and Cybernetics B 34 (1)*, pp. 95-109 (2004).

[28] Denœux, T., Sriboonchitta, S., Kanjanatarakul, O.: Evidential clustering of large dissimilarity data. *Knowledge-based Systems 106*, pp. 179-195 (2016).

[29] Ben Ayed, S., Elouedi, Z., Lefevre, E.: CEVM: Constrained Evidential Vocabulary Maintenance Policy for CBR Systems. *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pp. 579-592 (2019).

[30] Masson, M. H., Denœux, T.: RECM: Relational evidential c-means algorithm. *Pattern Recognition Letters, 30(11)*, pp. 1015-1026 (2009).

[31] Arshadi, N., Jurisica, I.: Feature Selection for Improving Case-Based Classifiers on High-Dimensional Data Sets. In *FLAIRS Conference*, pp. 99-104 (2005).

[32] Zhu, G., Hu, J., Qi, J., Ma, J., Peng, Y.: An integrated feature selection and cluster analysis techniques for case-based reasoning. In *Engineering Applications of Artificial Intelligence*, pp. 14-22 (2015).

[33] Kononenko,I.: Estimating attributes: analysis and extensions of RELIEF. In *European conference on machine learning*, pp. 171-182 (1994).

[34] Hong, T., Liou, Y.: Attribute clustering in high dimensional feature spaces. In *International Conference on Machine Learning and Cybernetics*, pp. 2286-2289 (2007).

[35] Maji, P.: Fuzzy-rough supervised attribute clustering algorithm and classification of microarray data. In *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, pp. 222-233 (2011).

[36] Pearson, K.: Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *In Philosophical Transactions of the Royal Society of London*, pp. 253-318 (1896).

[37] Masson, M. H., Denœux, T.: ECM: an evidential version of the fuzzy c-means algorithm. P*attern Recognition 41 (4)*, pp. 1384-1397 (2008).

The Impact of Data Analytics in Digital Agriculture: A Review.

1st Nabila Chergui
Faculty of Technology, Ferhat Abbas University, Setif 1
MISC Laboratory Abdel Hamid Mehri University, Constantine 2
Algeria
nabila.chergui@umc.edu.dz

2nd M-Tahar Kechadi
School of Computer Science
University College Dublin
Dublin, Ireland
tahar.kechadi@ucd.ie

3rd Michael McDonnell
School of Business
University College Dublin
Dublin, Ireland
michael@ucd.ie

***Abstract*—The recent advances in Information and Communication Technologies (ICT) have significant impact on all sectors of the economy worldwide. Digital Agriculture appeared as a consequence of the democratisation of digital devices and advances in artificial intelligence and data science. Digital agriculture created new processes for making farming more productive, efficient, while respecting the environment. Recent and sophisticated digital devices and data science allowed the collection and analysis of vast amounts of agricultural datasets to help farmers, agronomists, and professionals understand better the farming tasks and make better decisions. In this study we focus on the techniques used to analyse agriculture data and their effectiveness, more precisely we present a systematic review of methods and techniques of (data & big data) mining and their applications to digital agriculture from the big data view point. We limit our study to crop yield and its monitoring. The major mining techniques used so far in digital agriculture are classification and prediction. We identified major categories of mining techniques for crop yield monitoring. This is followed by discussing each category of the classification throughout a panoply of existing works and show their used techniques, then we provided a general discussion on the applicability of big data analytics into the field of digital agriculture.**

***Index Terms*—Data Mining, Data Analytics, Big Data, Machine Learning, Digital Agriculture, Crop Yield Monitoring.**

I. Introduction

The increasing demand for improving the productivity of both small and large farms by reducing resource costs, such as water, fertilisers and pesticides, requires the use of new advanced farming and management techniques. On the other hand, the improvement of productivity does not have to be in the cost of decreasing the quality of products which will harm the health or create damages to the environment, because of the excess use of fertilisers and pesticides and other agricultural inputs. Digital Agriculture (DA), (or digital/smart farming) [1]–[3], is an advanced approach that makes farms and the act of farming smarter by integrating the use of digital and smart tools like (sensors, cameras, satellite, drones, GPS, etc.) in conjunction with AI techniques, Internet of Things (IoT), data mining and data analytics to improve the agricultural practices, to enhance the productivity and to optimise the use of resources by providing insights and decision-making supports to farmers. DA can for example controls a crop nutrition by finding the optimum fertilisation program for each field, its optimum irrigation program, and can help farmers to react differently for each part of the field.

The DA involves the development, adoption and iteration of all above-mentioned technologies in the agricultural sector in different spatial contexts [4].

DA is data-driven solutions, which across the use of ICT, AI and Data analytics, permits to farmers to adopt these solutions for their businesses. These solutions can vary and cover multiple activities of farming, including assessing risks and disasters, producing predictive models and so on.

We can summarise the benefits of DA to agriculture in:

- It provides the farmer with useful information and supports their decision making with regards to how much, when and where to apply nutrients, water, seeds, fertilisers, and other chemicals and agricultural inputs.
- By varying the amount of growth resources (fertilisers, pesticides, and irrigation) used for crop production, and applying those inputs with exact quantities in each field, the environment is sustained [5], and healthy products are guaranteed.
- Data-driven enable farmers to access to sophisticated management solutions against climate change and other environmental challenges and natural events. Farmers can continuously monitor crop health, and predictive analytics can even alert farmers to likely problems with pests or disease or even climate change.
- From the marketing view, farmers can also benefit from advanced models that give insights on the market and which products could bring more profits to them.

In the past, the full potential of DA was not possible. Nowadays, data gathering and data mining & analytics techniques are commonplace in every sector of the world economy. This was made possible with technological advances (sensor devices, satellite images, advanced weather stations, etc.) and also advances in digital devices and the ability to store and process nearly everything. Today we can store vast amounts of data. Besides, with the use of advanced data mining techniques, we can extract novel and useful knowledge from these large volumes of historical datasets, which can help us understand the behaviour of both crops and farmed fields and, therefore, use efficient management techniques of the whole

farming industry, such as field and wasteland management, crop and pest management, soil classification, etc.

Crop management is a key task in DA, as it impacts directly on the crop production; it assists crops from soil preparation and seed selection, watering, and so on, until the harvest day, and can go beyond post-harvest. It offers the most intensive measure yield variability that exists in farm fields, allowing producers to assess how management skills and environmental factors affect crop production [5]. This assessment provides direct and valuable feedback to farmers enabling them to make better decisions [6] at real-time and monitor the farm proactively. The crop management process, if we ignore the marketing of products, can be devised into five sub-processes as described below, and as presented in in Figure 1.

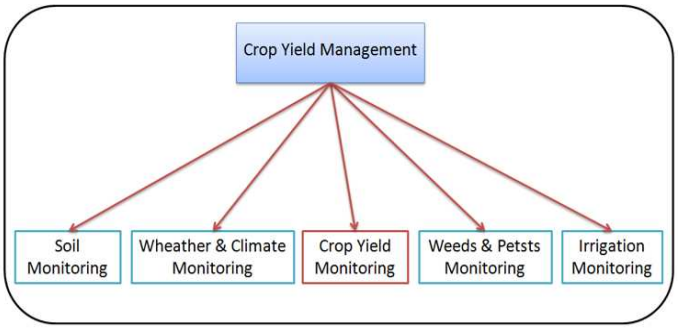


Fig. 1. The crop yield management components.

- **Soil monitoring:** it aims at studying the nature of the soil to select the type of culture to plant. Also, it controls the concentration of fertiliser and, hence, the quantities to use depending on the soil type, and it controls the irrigation operation based on the soil moisturising.
- **Weather & climate monitoring:** This deals with the monitoring of temperature, wind and other environmental factors that can affect the crop.
- **Weed and Pest monitoring:** it controls the insects and weeds that can affect each type of crop, and the type and quantities of pesticides and herbicides that should be applied at a given location of the farm;
- **Crop Yield monitoring:** it aims at monitoring crop conditions during the growing season, the estimation of the crop yield, crop quality assessment methods, detection and protection of crops from diseases, the delineation of management zones of yields, and all the other related crop operations.
- **Irrigation monitoring:** it takes into consideration the amount of water needed by each type of crop and the irrigation rate, which also depends on weather conditions, the season, the soil type, and the crop's growth cycle.

Vast amounts of data were gathered from each of these processes. Its exploitation using the potential power of data analytic techniques will offer a solid decision-making support to farmers and will have a significant impact on crop production and on the environment conservation. Data mining has several applications in crop yield management, such as the understanding of vegetation variables, soil mapping and classification, zone management, weather forecasting, disease protection, prediction of the market crop prices, rainfall forecasting, weed detection and yield prediction, etc. In this work we will focus on the crop yield monitoring, which is part of the management process.

It is clear, and mainly in DA, that more data we collect more insights we can acquire from it and more accurate the results will be. Nowadays, collecting data is not a hurdle anymore. We live of the era of big data and IoT, and very large amounts of data can be collected from various sources. For instance, data can be originated from crop yield, patterns and rotations, weather, climate and environmental conditions and parameters, soil types, moisturising and nutrients, and from farmers' records on yields and other factors. This collected data is not only big, but also heterogeneous in types and quality. Therefore, its analysis is very challenging.

Part of this data heterogeneity in DA came from the way the data were collected, as each data collection technique has different characteristics in its accuracy, validity, and impact on farmland.

Accordingly, another part of this heterogeneity caused by the type of the used devices to collect data, different sensors, different records and different cameras, etc.

Considering these facts, and in light of its source and nature, data can belong to one of the following type classes: historical data, sensor data, image data or satellite data.

The reminder of the paper is organised as follows: Section II presents related works on the application of data mining & analytics, AI and machine learning to crop monitoring. In addition it describes a classification of data mining techniques applied for crop yield monitoring. Section III discusses classification techniques for crop yield. Section IV shows the techniques used for the prediction of crop yield for both types of data. Section V exhibits the techniques used to protect crops from diseases, pests and weeds. Section VI presents techniques used to detect crops and estimate yields. Section VII is dedicated to clustering techniques used for crop yield. The evaluation of several existing works is presented in VIII. Finally, we conclude the work in Section IX.

II. Crop Yield Management

Various studies have been conducted on the application of data mining & analytics to crop yield management. For instance, [7] discusses the forecasting yield by integrating agrarian factors and machine learning models. [8] provided a systematic review on the use of computer vision and AI in DA for grain crops.

[9] reviewed the use of big data analysis in some fields of agriculture. The authors concluded that the use of big data analytics in agriculture is still at its early stage and many barriers need to be overcome despite the availability of the data and tools to analyse it.

[10] presented a review of advanced machine learning methods for detecting biotic stress in crop protection.

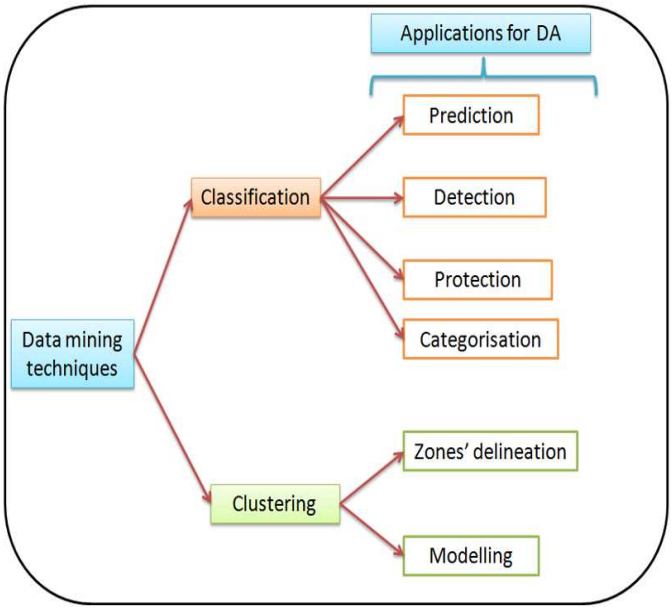


Fig. 2. Data mining techniques applied for crop yield monitoring.

An early similar study was presented in [11], where the authors studied four very popular learning approaches in the area of agriculture; Artificial Neural Network (ANN), Support Vector Machine (SVM), K-means and K-Nearest Neighbour (KNN).

The study presented in this paper is not just an update about what has been done in the previous surveys. As big data is commonplace nowadays, the first objective is to examine the application of big data analytics to DA and more specifically on crop yield monitoring. The second objective is to discuss how it is applied and what are the encountered challenges.

So that, the motivation behind the preparation of this review is to figure out how much the big data is employed in DA, and whether the application of data mining techniques in DA implies the use of big data too. Besides, if applicable, how big data and data analytics are leveraged to the benefice of DA.

Thus, the main contribution of this study is that it presents a deeper analysis of problems encountered in agriculture and resolved by DA, and more focused overview of an important and particular problem, the crop monitoring, compared to the above-mentioned surveys. Furthermore, our study highlights the type of data used, the methods and techniques employed and for which class of data mining techniques.

From the analysed literature, the application of data (mining/analytics) techniques to crop yield monitoring is almost limited to classification and clustering. Figure 2 proposes a classification of data mining techniques applied to crop yield.

Based on the type of data, data mining process can use various pre-processing techniques before starting the analysis of data. For instance, for image-based data, we can use the process described in Figure 3. Once data have been cleaned and pre-processed, the resulting data should be of high quality and ready for the analysis. Depending on the question to be answered and whether the historical data was annotated

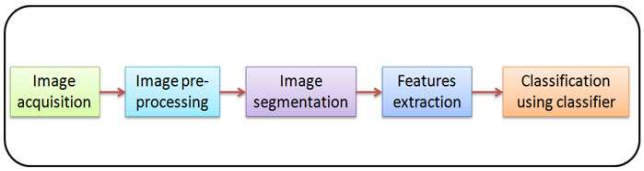


Fig. 3. Image processing approach.

or not, we choose the category of the analysis techniques (classification, clustering, etc.). Classification techniques, for example, are used for the purposes of prediction, detection, protection, and categorisation tasks, which are the most important tasks involved in the crop monitoring process. The following sections discuss the research that has been conducted in each of the four principal tasks of the crop monitoring process.

III. Crop Classification

Many mining approaches have been used based on both collected datasets and the target objective [12]–[17]. In this section we review some of these works. To identify and classify potato plants and three common types of weeds, [12] used a machine vision system, which consists of two subsystems: a video processing subsystem that is capable of detecting green plants in each frame; and a hybrid approach that combines artificial neural network (ANN) and particle swarm optimisation algorithm (PSO) to classify weeds from potato plants. The PSO is used to optimise the ANN's parameters and the ANN for classification. The hybrid approach was compared to Bayesian classifier. The experimental results show that ANN-PSO and Bayesian Network (BC) achieved an accuracy of 99.0% and 71.7%, respectively, on the training set, and 98.1% and 73.3%, respectively, on the test dataset.

[13] used a multilevel deep learning architecture to classify land-cover and crop based on multi-temporal multi-source satellite imagery data. They pre-processed the data by segmenting the imagery data and restoring the missing data due to clouds and shadows before classifying it. The proposed hybrid approach was compared to Random Forest (RF) and Multi-layer Perceptron (MLP). They showed that their approach outperforms both RF and MLP and obtained better discrimination of certain summer crop types, such as maize and soybeans, with an accuracy more than 85% for all major crops (wheat, maize, sunflower, soybeans, and sugar beet).

Deep learning approach has been also used for plants species and weeds classification, based on coloured images issued from six different data sources [14]. They used Convolution Neural Network (CNN)on a dataset consisting of $10,413$ images with 22 weeds and crop species. The CNN model was able to achieve an accuracy of 86.2%. Both these techniques do not have high accuracy, as they misclassify certain crops.

[15] developed a hybrid classifier for four crops: corn, soybean, cotton, and rice, based on satellite images. The

approach is an ensemble learner that consists of an ANN, support vector machine (SVM) and decision tree (DT), and a combiner to generate a prospective decision. The overall approach generates a recommendation based on both the learners outputs and the available expert knowledge.

An SVM-based classifier for distinguishing crops from weeds based on digital images was suggested in [18]. It achieved an accuracy of 97%.

IV. Crop Yield Prediction

The estimation of crop yield aims to study factors that influence and affect the production, such as weather, natural soil fertility and physical structure, topography, crop stress, irrigation practices, incidence of pests and diseases, etc. It enables efficient planning of resources; an early and accurate prediction of yields can help decision makers to estimate how much to import in the case of shortage or export in the case of a surplus. In the following we discuss some prediction techniques and showing type of the datasets that were used for the prediction operation.

A. Soil and Weather Data

Crop yield prediction using historical data and time series has been studied for many years, an it is considered among the classical applications of data mining. In 1994, [19] used a fuzzy logic expert system to predict corn yield. The model obtained promising results. A year later, [20] used a feed-forward back-propagation NN to predict corn and soybean yields. The dataset used was based on soil properties, such as phosphorus, potassium, pH, organic matter, topsoil depth, magnesium saturation. Other factors, such as weather were not considered. The NN showed promised results as an aid in understanding yield variability, but its accuracy is not reasonably good. To improve its accuracy, [21] proposed a Multiple Linear Regression (MLR), projection pursuit regression (PPR) and several types of supervised feed-forward NN methods for site-specific yield prediction to study the relationships between yield and soil properties and topographic characteristics. The authors added a second phase of experiments to include climatological data. The NN techniques consistently outperformed both MLR and PPR and provided minimal prediction errors in every site-year. Besides, the results showed that a significant over-fitting had occurred and indicated that much larger number of climatological site-years would be required in this type of analysis.

Moreover, to analyse the weather aberrations impacting on rice production in mountainous region of Fujian province of China, [22] used a NN model. The historical dataset was collected from 16 locations throughout the region. The weather variables include daily sunshine hours, daily solar radiation, daily temperature sum and daily wind speed, in addition to the seven different soil types for each location. It was shown that the model is more effective compared to a multiple linear regression model.

MLP were employed for winter wheat prediction by [23], then two different neural networks are considered in [24], where it conducted a comparison of four regression models for yield prediction based on agricultural data yield obtained from a farm in Germany. Networks with MLP and RBF, The Support Vector Regression (SVR) and decision regression tree were implemented. The study showed that the SVR technique was the most suitable for this kind of problem.

[25] Demonstrated the applicability of RF to estimate the yield of mango fruit in response to water supply under four different irrigation regimes. A set of four RF models with different input variables: rainfall model, irrigation model, rainfall and irrigation model, and total water supply model was developed to estimate the minimum, mean and maximum values for each of the mango fruit yields, namely 'total yield' and 'number of marketable mango fruit', using a combination of 10 days rainfalls and irrigation data.

[26] Presented a scalable machine learning system for pre-season agriculture yield forecast. It employed a Recurrent Neural Network (RNN) fostered by data from multiple resources: satellite-derived precipitation data, soil properties data sets, seasonal climate forecasting data from physical models and historically observed soybean yield to produce a pre-season prediction of soybean/maize yield for Brazil and USA. The results showed the error metrics for soybean and maize yield forecasts are comparable to similar to the other systems.

[27] Conducted a study to compare the predictive accuracy of several machine learning methods (MLR, M5-Prime regression trees, MLP, SVR and K-nearest neighbour) for crop yield prediction in ten crop data sets. It used data on solar radiation, rainfall, temperature, season-duration cultivar, planting area, etc. It concluded that the M5-Prime is a very suitable tool for massive crop yield prediction in agricultural planning.

[28] Evaluated the ability of RF to predict crop yield regarding climate and biophysical variables at global and regional scales in wheat, maize and potato in comparison with MLR which was served as a benchmark. It employed crop yield data from various sources and regions in the USA over 30 years for model training and testing. The results demonstrated that RF was highly capable of predicting crop yields and outperformed MLR benchmarks in all performance statistics that were compared. The Extreme Learning Machine (ELM) has been employed by [29] to the estimation of the Robusta coffee yield based on soil fertility properties. The performance of 18 different ELM-based models was evaluated with single and multiple combinations of the predictor variables based on the soil organic matter (SOM). The MLR and RF have been chosen for the comparison, the results indicated that the EML outperformed the MLR and RF models.

B. Image-based Datasets

[30] Explored the value of combining data over multiple sources and years into one data-set in conjunction with machine learning approaches to build predictive models for pre-sowing, mid-season and late-season crop yield. Yield from wheat, barley and canola crops from 3 different seasons. The collected data consisted of yield data, the electromagnetic

induction survey EM and gamma radiometric survey, Normalised Difference Vegetation Index (NDVI) and rainfall. RF models were used to predict crops yields using the space-time cube. The models performed better as the season progressed, because more information about within-season data became available (e.g. rainfall).

[31] Built a model for estimation of citrus yield from airborne hyper-spectral images using an ANN. It used an airborne imaging spectrometer for applications eagle system to acquire images over a citrus orchard. The work concluded that its obtained results demonstrated that the ANN model could work well for the observed hyper-spectral data, and suggested potential of using airborne hyper-spectral remote sensing to predict citrus yield.

[32] Proposed a wheat yield prediction using machine learning and advanced sensing techniques. The aim of this work is to predict within-field variation in wheat yield, based on on-line multi-layer soil data, and satellite imagery crop growth characteristics. The performance of CP-ANNs, XY-fused Networks (XY-Fs) and Supervised Kohonen Networks (SKN) for predicting wheat yield were compared for a single cropping season. Results showed that the accuracy of the SKN model and its performances outperform the two other models.

[33] Established a yield prediction model for cabbage using the Green-seeker hand-held optical sensor that provides useful data to monitor plant status, and the NDVI. The NDVI measurements are commonly used for characterising the nitrogen condition or biomass development of plants, and other nutritional monitoring in field crops of elements like potassium and phosphorus. The modified exponential, linear and quadratic functions were used to regress cabbage yields with NDVI measurements collected during growing seasons.By comparison, the exponential equation showed a better performance than the linear and quadratic functions.

[34] Developed two ANN models for early yield prediction of 'Gala' apple trees. These models were analysed 50 RGB images of the trees to identify the fruit. These two models were proved to predict yield accurately.

[35] Extended the model of [36] to predict soybean crop yields in Argentina and transfer the learning to Brazil using deep learning techniques (Long/Short Term Memory Network(LSTM)), and the Moderate Resolution Imaging Spectro-radiometer (MODIS) satellite imagery as data. To transfer learning from Argentina to Brazil, it initialised the LSTM model with the parameters from a neural network trained on Argentine soybean harvests. Then, it stripped out the last dense layer of the pre-trained model and replace it with an untrained dense layer of the same dimensions before training the modified model on the available Brazilian training data. The results in Argentina and Brazil demonstrated that this approach can successfully learn effective features from raw data.

[37] Used the MODIS EVI (Enhanced Vegitation Index) product combined with ground temperature measurements to predict corn yield using SVM and Deep Neural Network (DNN). The DNN is able to provide accurate predictions. In the same way, using MODIS EVI, NDVI, Leaf Area Index (LAI) and land cover together with climate data (precipitation and temperature).

V. Crop and Plant Protection

Forewarning, protection and detection of crop diseases correctly and timely when they first appear is a very important task of crop monitoring, it will reduce yield losses and inform and prevent farmers to take effective preventive actions. For detecting crop and plants diseases, several works have used image processing, consequently image-based data and classification techniques for detecting crops diseases' [38]–[42]. The general approaches of these works are almost similar, as described in the process presented in Figure 3. The approaches started by capturing and collecting images for disease plants using cameras, scanners or other sensors. After that, segmentation of plant disease spots, followed by an extraction of features like colour, shape or texture. In the end, the employment of classification methods, such as ANN, BC method, KNN, SVM, to classify disease images.

[38] Proposed a deep learning approach for crops classification and disease detection, based on image data-set of 54,306 images of diseased and healthy plant leaves collected under controlled conditions. It trained a deep CNN to identify 14 crop species and 26 diseases (or absence thereof). The results obtained demonstrate the feasibility of this approach.

Similarly, [43] adopted deep CNN to develop a model of plant disease recognition using leaf images database containing 30880 images for training and 2589 images for validation. The proposed model was able to recognise 13 different types of plant disease from healthy leaves and to distinguish plants from their surroundings. It achieved an average accuracy of 96.3% on the experimental analysis.

[39] Introduced an automatic detection and classification method for crop disease using plant leaf images. This method composed of four phases; it started by the image acquisition, then a pre-processing by creating of a colour transformation structure for the RGB leaf image followed by the application of device-dependent colour space transformation for the colour transformation structure. After that, segmentation of images performed by K-means clustering technique to calculate the texture features. After that, a classification of the extracted features using ANN.

For accurate and early detection of rice crop disease; [40] presented an application of SVM for detecting rice diseases spots. The results showed that SVM could effectively detect and classify these disease spots.

[44] Applied an ANN to discriminate fungal infection levels in rice panicles, using hyper-spectral reflectance and principal components analysis. Hyper-spectral reflectance of rice panicles was measured through the wavelength with a portable spectro-radiometer in the laboratory. A Learning Vector Quantization (LVQ) neural network which is a supervised learning technique that can classify input vectors based on vector quantisation has been used to classify fungal infection

levels into one class of (healthy, light, moderate, and serious). The results showed a good accuracy.

[45] Investigated the performance of four classification algorithms applied to the problem of classification of Egyptian rice diseases using historical data. In this study, a comprehensive comparative analysis of four different classification algorithms and their performance has been evaluated, namely: J-48 DT, Naive Bayes net, random trees (RT) and RF. The experimental results indicated that the J-48 DT achieved highest sensitivity, specificity and accuracy and lowest error, thence, gave the best results, where the Naive Bayes was the worst.

Deep learning has been used also by [46]–[50].

Among those, [46] used CNN model to perform plant disease detection and diagnosis using simple leaves images of healthy and diseased plants, from database incorporates 87,848 images, containing 25 different plants in a set of 58 distinct classes of (plant, disease) combinations, including healthy plants. It tried different model architectures for training, with the best performance reaching an 99.53% success rate in identifying the corresponding (plant, disease) combination (or healthy plant).

Another work [48] is proposed to identify various tomato diseases using a combination of super-resolution and conventional images to enhance the spatial resolution of diseased images and to recover detailed appearances, based on database of 18,149 images. A super-resolution CNN was used for super-resolution and it outpaced other conventional disease classification methods.

Five different architectures of deep CNNs have been evaluated by [49] for image-based plant disease classification. The architectures evaluated including: VGG 16, Inception V4, ResNet with 50, 101 and 152 layers and DenseNets with 121 layers. Openly and freely data-set from PlantVillage were used for this study, this data-set has 54,306 images, with 26 diseases for 14 crop plants. The results showed that the DenseNets had tendency's to consistently improve in accuracy with the growing number of epochs.

[51] Used SVM algorithm with NDVI and raw data inputs to develop weed-crop discrimination. The performance of this model was evaluated and compared with a conventional plant discrimination algorithm based on the measurement of discrete NDVIs and the use of data aggregation. Results showed that the use of the Gaussian-kernel SVM method in conjunction with either raw reflected intensity or NDVI values as inputs, provides better discrimination accuracy than that attained using the discrete NDVI-based aggregation algorithm. Another work, [52] employed the RF for recognising weeds in a maize crop using near-infrared snapshot mosaic hyper-spectral imagery. Experiments were conducted using three different combinations of features and compared with the KNN. Results presented an overall better performance of the optimal random forest model.

VI. Crop and Fruit Detection

Crop and fruits detection is a kind of crop prediction, but it is based on the detection of the presence of fruits from images. It aims to provide information to growers to optimise economic benefits and plan their agricultural work, in addition, to adjust management practices before harvesting and to decide proper investments in advance because it offers an early estimation of yields and fruit growth. The study presented in [53] investigated the possibility of using a deep learning algorithm and CNN for recognising two classes (mature and immature strawberry) based on greenhouse images. The study tried to propose solution for training and learning a CNN using a small set of data, it uses 373 images for training and for testing. The developed CNN achieved a good precision.

[54] Proposed an automatic, efficient and low-cost fruit count method of coffee branches using computer vision. The method calculated the coffee fruits in three categories: harvestable, not harvestable, and fruits with disregarded maturation stage, and estimated the weight and the maturation percentage of the coffee fruits. After the process of image pre-processing, three classifiers were implemented to perform tasks of the detection, classification and fruits' count; Bayes classifier; KNN; and SVM classifier. After that, and according to the experimental results, the authors have selected the SVM to validate their model because it outperformed the Bayes and KNN.

[55] Suggested a machine vision system for detecting cherry tree branches in planar architecture for automated sweet-cherry harvesting. The system was developed to segment and detect cherry tree branches with full foliage when only intermittent segments of branches were visible. A BC was used to classify image pixels into four classes: branch, cherry, leaf and background. The algorithm achieved good ccuracy in identifying branch pixels.

[56] Presented a method for the detection of tomatoes based on EM and remotely sensed RGB images. Images were captured by an unmanned aerial vehicle UAV. It employed several techniques of data mining, it first started by clustering technique using Bayesian information criterion to determine the optimal number of clusters for the image. Then, it used K-means to carry out the spectral clustering, where the EM and Self Organising Map (SOM) algorithms are utilised to categorise the pixels into two groups i.e. tomatoes and non-tomatoes. It is observed that EM performed better than K-means and SOM.

[57] Developed an early yield mapping system for the detection of immature green citrus in a citrus grove under outdoor conditions based on machine vision. As all other relative studies, and after images pre-processing and features extractions, SVM was applied to detect the immature citrus. The accuracy was achieved a good value.

Another model, called DeepFruits, was proposed in [58] for sweet pepper detection from imagery data, RGB colour and Near-Infrared. This work employed a CNN and adopted for the Faster Region-based CNN (Faster R-CNN). The model

was trained, and its performance through the measure of precision and recall was promising. [59] Presented computer vision algorithms for immature peach detection and counting in colour images acquired in natural illumination conditions using several classifiers and ANN. It used seven different classifiers, including some parametric and non-parametric ones: discriminant analysis for classification, Naive Bayes classifier, KNN classifier, classification trees classifier, regression trees classifier, SVM and ANN. It has concluded that considering overall performance of the classifiers, the parametricity was not a significant factor in detection performance with respect to classifier type. For example, the ANN classifier which is non-parametric and the DA classifier which is parametric provided successful detection in the experiments. While the SVM and the KNN classifiers which are non-parametric yielded good success rates with just one method, their detection accuracy was poor for the two other methods. However, the lowest detection performance was obtained by the Naive Bayes classifier for the three methods used in the experiments; this is because it is known that this kind of classifier is usually less accurate than other supervised methods due to its inadequate ability to deal with complex interactions between features.

VII. Clustering Techniques for Crop Yield

Clustering techniques are not widely employed in DA, few efforts have investigated the potential of such techniques. Although, there are two sub-classes of clustering for crop yields: modelling and delineation of management zones.

A. Modelling

[60] Proposed a methodology and life cycle model for data mining and knowledge discovery called KDLC Knowledge Discovery Life Cycle. The methodology consisted of 6 activities that guide and assist the user throughout the KDD process, and combined supervised inductive rule learning and unsupervised Bayesian classification via constructive induction mechanism to construct a multi-strategy knowledge discovery approach. This approach started by analysing data sets using an unsupervised Bayesian classification system to discover interesting taxonomic classes, and these can be represented as new attributes in an expanded representation space via constructive induction mechanism, this later was then used to learn useful concepts, relationships, and rules that characterise knowledge in the data space. The case study dealt with crop yields for a farm in the state of Idaho to define the highest yield regions.

[61], [62] Introduced a modelling approach of Fuzzy Cognitive Map (FCoM) to help on the decision-making. it utilised FCoM learning algorithms to handle initial knowledge. It used the soft computing technique of FCoMs which enriched with an application of unsupervised learning algorithm (the nonlinear Hebbian learning (NHL)), to characterise the data into two production yield categories, and to describe the cotton yield management in smart farming, especially, the estimation of yield trend is a complex process. The aim of this work is to present a methodology that can determine cotton yield behaviour in smart farming, based on artificial intelligence techniques and particularly based on aspects related to knowledge representation. The proposed methods are dependent on the group of experts who operate, monitor, supervise the system and they know its behaviour. This methodology extracted the knowledge from the experts and exploited their experience of the system's model and behaviour.

The NHL algorithm is proposed to train FCoM. It is introduced to overcome inadequate knowledge of experts and/or non-acceptable FCoM simulation results and to adapt weights. The accuracy of the FCoM model and the implementation of the NHL algorithm for 'low' and 'high' yield categories for the three respective years of 2001, 2003 and 2006 has achieved an acceptable success rates.

[63] Presented a hierarchical grading method applied to JonaGold apples. The unsupervised learning K-means clustering algorithm was used, and the number of clusters K was fixed by preliminary studies to 16 for the variety of apples. A principal component analysis was carried out and the K first principal component representing 97% of the whole variations were used to compute a quadratic discriminant analysis to finally grade the fruits. The global correct classification achieved acceptable rate.

[64] Proposed intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from colour images. The clustering algorithms: fuzy C-Means (FCM) and Gustafson-Kessel (GK) in conjunction with Fuzzy excess red (ExR) and excess green (ExG) indices were used for unsupervised classification of hidden and prominent regions of interest (ROI) in colour images, included sunflower, redroot pigweed, soybean, and velvetleaf plants, against bare clay soil, corn residue and wheat residue, typical of the Great Plains. Clusters and indices were enhanced with Zadeh's (Z) fuzzy intensification technique; so that, they have used a ZFCM and ZGK algorithms, in addition to ZExG index and ZExR index. The paper concluded that the ZGK algorithm could be potentially useful for remote sensing, mapping, crop management, weed, and pest control for precision agriculture.

B. Delineation of management zones

Usually, farmers split their agricultural land into fields for several reasons; to variate their crops and make crop-rotation practices, to facilitate the management tasks and to create their yields maps to help them in enhancing their crop yields. This process is called delineation of management zones (DMZ). DMZ of yields is an important task for crop monitoring since it aims at determining zones of low-or-high yields, and to find out reasons behind low yield fields, hence, to propose solutions to increase yields of the fields known by their low productivity.

From the DA view, given an area divided into zones, DMZs process tries to find adjacent zones which exhibit similar characteristics or homogeneous with other zones, at the same time are heterogeneous and have different characteristics with other zones. This can be translated using data mining vocabulary by clustering. Based on the literature review, and for both techniques for delineation of yields' zone management, the

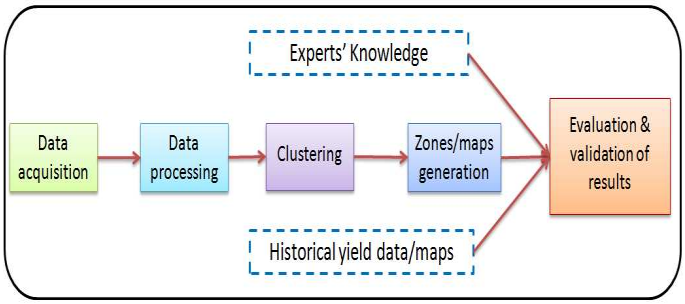


Fig. 4. The delineation management zones process.

most used technique is k-means clustering (non-hierarchical method).

Figure 4 depicts the general process of delineation of management zones followed by the majority of proposed works.

[65] Presented a study that aimed to point out the possibility of using the hierarchical method as complementary to the non-hierarchical clustering method in order to estimate a statistically significant number of management zones of a given field which can be used as an input for the non-hierarchical method. For this end, it collected data from yield monitoring during three seasons for three different crops (Spring barley; oil-seed rape; winter wheat). The cubic clustering criterion (CCC) has been used to estimate the statistically significant number of clusters for both Ward's method (hierarchic method) and k-means method. The conducted clustering uses in the first step the hierarchic method (Ward's method) in order to explore the data based on the homogeneity which allows to identify the homogeneous places, followed by the non-hierarchic method (k-means clustering) which creates clusters which are as heterogeneous as possible. The results of this study showed that it was beneficial to use both the hierarchical and non-hierarchical clustering methods when determining the management zones from yield maps.

Using the same above mentioned method, [66] proposed to use it in order to evaluate the potential of identifying the yield potential zones based on historical yield maps, and to evaluate the procedure over the growing extent of input data. It used historical data for yields maps from six growing seasons (spring barley, winter wheat, spring barley, spring oil-seed rape, winter wheat, spring barley) and designed 67 monitoring points. Results showed that using these data from commercial combine monitoring systems enables determining the zones, despite its complexity. Multiple yield data are recommended as the values of analyses increase with the increased number of input data sets.

K-Means algorithm groups data by the similarity of their values, using some distance measures such as the Euclidean distance, but for the task of delineation, data are also characterised by geographical coordinates associated with each sample, which leads to the idea of grouping data using other factors like the geographical distance determined between samples. However, [67] data are associated with a location which causes an inherent degree of vagueness because of several factors such as the accuracy of the GPS capture device. Moreover, making the groupings not reliable if obtained as hard clusters, as by the k-means algorithm. For this reason, clustering algorithms that take into account the data imprecision, like the fuzzy clustering, and allow a sample to belong to more than one cluster with a certain degree of membership, should be used for applications in PA. One of the most widely used fuzzy clustering algorithms is the FCM) algorithm.

[68] Proposed an approach for the delineation of Potential Management Zones (PMZ) for differential crop management that expresses the productive potential of the soil within a field, using farmer's expert knowledge and data on yields. It used yield maps, remote sensing multi-spectral indices, apparent soil electrical conductivity, and topography data for their cluster analysis. It implemented FCM to create different alternatives of PMZ, and then the farmers' expert knowledge was taken into account to improve the resulting PMZs that best fitted to the yield spatial variability pattern. The validation of PMZ was done with yield data in maize (Zea mays L.) field acquired at the end of the season.

In the same way, numerous works used either KMeans or FCMs for DMZs of yields; among others, [69] proposed a management zones delineation using fuzzy clustering techniques in grapevines, using Soil properties, yield and grape composition data. [70] Applied FCMs technique in a geo-referenced yield and grain moisture data set in order to find the optimal number for homogeneous zones. The best results presented by this algorithm ranged from 8 to 10 zones which were validated using the indexes Partition Coefficient, Classification Entropy and Dunn's Index. [71] Used k-means cluster analysis, a multivariate analysis of variance and discriminant analysis in an attempt to evaluate a framework for delineating PMZs in cotton. [72] Developed software called ZoneMAP to automatically delineate MZs using satellite imagery and field data provided by users, processed with FCM algorithm. [73] Used FCM to delineate management zones considering historical yield data from corn-soybean rotation crops, identifying the spatial association of the obtained maps with soil maps. It proposed cluster analysis of yield and soil properties as the basis for delineating MZs. [74] Used FCM clustering to delineate MZs on NDVI, salinity and yield data in cotton. It provided an effective decision making support tool for variable-rate applications.

VIII. Discussion

In this section, we will draw some remarks regarding the application of data (mining/analytics) in DA and their extent of use of big data concepts.

From the data mining perspectives, methods analysed in this review, regardless the examined discipline as described in Figure 2, employ the same process: data/image acquisition, then processing and features selection, followed by the clustering or classification task; which is considered as a typical data mining application. Of course, each discipline has its specificities, and each problem within the same discipline has

its characteristics, obstacles, and has a special kind of data with variation in type and volume, thence each problem requires a particular solution and adequate algorithm considering the above-mentioned differences.

Therefore, it is not wise to recommend a solution or an algorithm for a problem, to prefer a solution than another or to judge an algorithm that it is better than another one. Although the application of machine learning algorithms in DA is intuitive, challenges encounter this application and the performance of these algorithms are non-trivial. It is well-known that machine learning algorithms are sensitive to data used to train them. As we have seen, computer vision and artificial intelligence have also emerged in DA in conjunction with machine learning.

Images, sensors and satellites data types require a pre-processing and segmentation before using them which has been performed using AI techniques. The quality of the result of image processing depends on the quality of images, which in its turn, depends on the acquisition step, the used devices to capture them, to the lightning background and other factors. It is obvious that these facts influence the performance of the machine learning algorithm. This reality is highly manifested in the case of delineation of management zones, where the process is almost the same, even for the choice of the clustering algorithm; the differences were in the choice of data type, choices for algorithms to process data, the number of features, etc. The same remarks are relevant to the other applications like crops classification, crops protection and detection.

In the case of yield prediction; we have found two kinds of forecasting depend on the nature of the used data:

- A yield forecast system based on historical data provides a pre-season estimation of the yield, and even before the beginning of the crop season which can give time and capability to farmers to make decision on which strategy to follow and which step to change to enhance the crop yield ahead in the crop cycle, like choosing seeds and crop type that match more the climate and weather variations, soil state, etc. In addition, it can work in large fields without any additional cost. The performance of machine learning algorithms for this type of forecasting depends on the quality-quantity of data which will feed the algorithms.
- On the other hand, forecasting system based on the other types of data regarding images issued by satellite, cameras, scanner, sensors, allows for on-season estimation, at the beginning of the growing season, at the middle or at the end of season just before the harvest, since it requires treating images for the crops upon which it performs the estimation. So that, it can provide too, near real-time insights into the state of crops and other problems like diseases. This system can work on small to large fields but with additional cost for images' acquisition. For example, it is known that satellite imagery is expensive and not within the reach of all farmers, in addition, most of the reviewed works showed that the use of the NDVI data is time-consuming to be acquired and processed. The performance of machine learning algorithms using these data depends on the quality of images, to the image processing and features selection process.

Another observation from the analysed works revealed that the ANN with its numerous implementations had dominated the prediction tasks, the SVM for the classification and detection, and the K-means for the clustering. It is worth to notice the emergent use of deep learning and CNN for the most recent works on detection, classification and prediction, which in fact presents impressive results. This is because of the availability of sophisticated computational hardware like GPU which allows for the processing of voluminous and complex data.

From the big data perspectives, an application is said to be applied big data concept if the used data-set can be described by four primary characteristics: volume, velocity, variety and veracity, commonly know by 4Vs.

- Volume (V1): The size of data collected for analysis;
- Velocity (V2): The frequency of collecting data. Data is accumulated in real-time and at a rapid pace, occasionally, yearly...;
- Variety (V3): The nature of data used and its sources: historical or image, or a combination of both. For example it can have a multi-sources from reports, images, videos, remote and sensing data...;
- Veracity (V4): The quality, reliability and the accuracy of the data;

It is obvious that more data is complex regarding its 4Vs more their analysis is complex too, so, considering the employment of the four metrics of big-data in DA we draw the following remarks:

- Velocity: it is remarked that many works do not mentioned the frequency of accumulating their data. Generally speaking, in DA the frequency of collecting data depends on the nature of data itself and to the problem for which data were collected. Some applications need a real-time data and others do not. Data-set for the prediction of crop yields used historical data have low velocity comparing to data-set collected for the protection and disease detection of crops using sensors or other type of image data which require a day per day control and hence, real-time data.
- Variety: most of the examined works used multi-source of data and in many cases a combination of historical and image data were exploited.
- Veracity: it is observed that most of the used data-sets needed to be cleaned and pre-processed, and that more the variety and velocity is high in the used data-set more its veracity is high too. [75] State that increased variety and high velocity hinder the ability to cleanse data before analysing it and making decisions, magnifying the issue of data 'trust'.

Considering the volume metric of big data, it is noticed that the volume of data-sets used by most of the analysed works does not meet the standard of big data. This reality is due to the following points which are considered are barriers against the full utilisation of big data in DA:

- DA is a new concept for farmers, and it is not applied until very recently. Before, farmers are not interested and motivated to collect data except for few cases like yields and weather conditions, or for statistics' purposes;
- Data-sets are usually collected from individual small farms and laboratories which cannot allow to generate a big mass of data.
- Big mass of data can be generated by big farms which usually belong to big companies, so that for security reasons and for competitions too, these companies avoid and do not prefer to share their data or to publish it.

Besides, the volume metric is highly depends on the nature of data, application employed satellite data for example is expected to use a high volume of data due to the size of pictures. It can depends too to the nature of the problem, size of data for yield prediction problems has tendency to be smaller than those used for crop protection. Moreover, with today sophisticated machines and algorithms, the volume of data is not the most important factor to worry about and it is not the most critical challenge. Veracity, velocity and variety are more essential and crucial because they add more complexity to the analysis and to the pre-processing of the data.

For these reasons, we are not considerate the volume criterion, which indicates the size of data-set and shows how much is it big. In addition, the bigness is not entirely about the size of data set, but also about the other three elements.

Table I, resumes a set of representative papers in DA according to their usage of big data. For each paper, we identify the type, the size, the heterogeneity of data used, and the frequency of its collection. In addition, we consider the number and type of machine learning algorithms used, the complexity of the proposed analyse algorithms and the used device to collect data.

From Table I, we can extract three classes of applications according to their usage intensity of the big data metrics: Full usage, light usage, non usage.

- Full usage: are applications that fully employ big data by all of its elements;
- Light usage: are applications that partially employ big data elements;
- Non usage: are applications that do not have any kind of use of big data concept and elements.

IX. Conclusion

Digital agriculture is in the way to re-shape the farming practices by making it more controllable and accurate; its key component is the use of information and communication technologies, sensors, GPS and other technologies for the benefit of farmers and the enhancement of its crops.

Data mining and machine learning techniques are reliable techniques for analysing data and exploring new information from these data. On the other hand, big data adds additional support to DA by discovering further insights from the collected data in order to solve farming problems and inform farming decisions.

This survey presented a systematic review of the application of data mining techniques and machine learning methods in the agricultural sector. It was first exhibited the process of crops management and its different parts, where we are focused on the crop yield monitoring. Then, for this later, it has provided a classification of the several employment of data mining techniques into this field. For each class of the classification, a set of existing works have been reviewed to demonstrate the machine learning method applied and for which purpose.

After that, the survey discussed the applicability of big data concepts, and it demonstrated that DA is on the road to exploit the full potential of big data concepts. This will open the gate to new opportunities of investment into these fields and will allow for a very different management way of crops, it promises new levels of scientific discovery and innovative solutions to more complex problems. In addition, it will provide farmers with new insights into how they can grow crops more efficiently.

The survey established that despite all the advantages gained from DA, there are several challenges and obstacles need to be surmounted in order to make from DA a real data-driven solution, among them lack of data because of several reasons like data ownership rights, data (or agricultural knowledge) providers needs guarantees for both their investments (money) in DA and for their security (competitions and many other facts), in addition, they usually need to acquire new skills to understand new technologies, which means an additional investment in term of time and effort.

To conclude, we say that "if energies are the soul of machines, then data are the spirit of algorithms".

TABLE I
DA APPLICATIONS AND THEIR USAGE OF BIG DATA CONCEPTS.

Ref	Volume	Velocity	Veracity	Variety	ML	Complexity	Device	Task
[18]	224 images	/	No	Image Data digital images	SVM	$O(n^2p + n^3) + O(n_{sv}p)$	Digital camera	Classification
[73]	3*2 years of data monitoring	1 year	/	Sensor data: soil properties	Fuzzy C-means	time: $O(ndc^2i)$ space:$O(nd + nc)$	Pressure-based: AgLeader Ames,IA	Clustering
[15]	/	/	No	Satellite data: Images in GeoTiff	EL (DT+ SVM+ ANN)	$O(n^2p) + O(p) + O(n^2p + n^3)$ $+O(n_{sv}p)+$ $O(epn(nl_1nl_2 + nl_2nl_3 + ...)+$ $O(pnl_1 + nl_1nl_2 + nl_2nl_3 + ...)$	Satellite	Classification
[20]	3000 for topographical data, 3120 data points for the other types	1 year for crop yield and soil's composite 1 day for climat	No	Image and sensor data: Soil properties Topographic crop yield climatological	ANN	$O(epn(nl_1nl_2 + nl_2nl_3 + ...)+$ $O(pnl_1 + nl_1nl_2 + nl_2nl_3 + ...)$	Camera Nikon Topgun A200LG electromagnetic induction sensor Yield sensor and GPS	Prediction
[30]	/	/	Yes	All types of data: yield, soil information Geo-physical Remote sensed Climate	RF	$O(n^2pn_{trees}) + O(pn_{trees})$	Yield monitor soil-maps, EM gamma survey MODIS NDVI	Prediction
[66]	229	1 year	Yes	Historical data: Crop yield	K-means	$O(ncdi)$	/	Clustering
[14]	10413	/	/	Image data: Digital images	CNN	$O(TQtq)$	Cell phone	Classifiation
[29]	/	1 year	No	Historical data: Crop yield soil parameters	ELM	$O(L^3 + L^2n)$	/	Prediction
[59]	96	1year	Yes	Image data: Digital images	SVM,ANN,NB KNN DT Discriminant analysis	Discriminant analysis: $O(np^2)$ NB: $O(np) + O(p)$	Camera Nikon CoolpixL22	Classification
[36]	8945	multi-spectral image: 8 days interval for 30 times a year	Yes	Satellite and sensor data: surface reflectance land surface temperature land cover	Gaussian CNN	$O(TQt2)$	MODIS satellite	Prediction

Where:

No: data were clear and all samples have been used;
Yes: data were cleaned and filtered and some samples were not considered because of abnormalities, inconsistencies or duplication and for other reasons;
n: number of training simple or data points;
n_{sv}: number of support vectors;
P: number of features;
n_{trees}: number of trees;
c : number of cluster;
d : number of dimension;

i : number of iterations;
L: number of hidden layers;
TQ is the size of input feature map; spatial, two/three-dimensional kernels are of size (tq);
nl_i :number of neurons at layer i;
ep: number of epochs.

REFERENCES

[1] K. Poppe, S. Wolfert, C. Verdouw, and T. Verwaart, "Information and communication technology as a driver for change in agri-food chains," EuroChoices 12, 60–65, Tech. Rep., 2013.

[2] K. Soma, M. Bogaardt, K. Poppe, S. Wolfert, G. Beers, D. Urdu, M. P. Kirova, C. Thurston, and C. M. Belles, "Research for agri committee - impacts of the digital economy on the food chain and the cap. policy department for structural and cohesion policies," European Parliament. Brussels, Tech. Rep., 2019.

[3] "Preparing for future akis in europe," European Commission. Brussels, Tech. Rep., 2019.

[4] S. Wolfert, C. Verdouw, and M. Bogaardt, "Big data in smart farming – a review," *Agricultural Systems*, vol. 153, pp. 69–80, 2017.

[5] T. Stombaugh and S. Shearer, "Equipment technologies for precision agriculture," *Journal of Soil and Water Conservation*, vol. 55, no. 1, pp. 6–11, 2000.

[6] G.Pelletier and S. Upadhyaya, "Development of a tomato load/yield monitor," *Computers and Electronics in Agriculture*, vol. 23, pp. 103–117, 1999.

[7] D. Elavarasan, D. Vincent, V. Sharma, A. Zomaya, and K. Srinivasan, "Forecasting yield by integrating agrarian factors and machine learning models: A survey," *Computers and Electronics in Agriculture*, vol. 155, pp. 257–282, 2018.

[8] D. Patricio and R. Rieder, "Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review," *Computers and Electronics in Agriculture*, vol. 153, pp. 69–81, 2018.

[9] A. Kamilaris, A. Kartakoullis, and F. Prenafeta-Boldu, "A review on the practice of big data analysis in agriculture," *Computers and Electronics in Agriculture*, vol. 143, pp. 23–37, 2017.

[10] J. Behmann, A.K.Mahlein, T.Rumpf, C.Romer, and L. Plumer, "A review of advanced machine learning methods for the detection of biotic stress in precision crop protection," *Journal of Precision Agriculture*, vol. 16, pp. 239–260, 2014.

[11] A. Mucherino, P. Papajorgji, and P. M. Pardalos, "A survey of data mining techniques applied to agriculture," *Journal of Operational Research*, vol. 9, no. 2, pp. 121–140, 2009.

[12] S. Sabzi and Y. Abbaspour-Gilandeh, "Using video processing to classify potato plant and three types of weed using hybrid of artificial neural network and particle swarm algorithm," *Measurement*, vol. 126, pp. 22–36, 2018.

[13] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, "Deep learning classification of land cover and crop types using remote sensing data," *Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 778–782, 2017.

[14] M. Dyrmann, H. Karstoft, and H. Midtiby, "Plant species classification using deep convolutional neural network," *Biosystems engineering*, vol. 151, pp. 72–80, 2016.

[15] S. Contiu and A. Groza, "Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning," *Expert Systems With Application*, vol. 64, pp. 269–286, 2016.

[16] A. Formaggio, M. Vieira, and C. Renno, "Object based image analysis (obia) and data mining (dm) in landsat time series for mapping soybean in intensive agricultural regions," in *Proceedings of IEEE International Geoscience and Remote Sensing Symposium.*, Munich Germany, Jul. 2012, pp. 2257–2260.

[17] J. Arribas, G. Sanches-Ferrero, G. Ruiz-Ruiz, and J. Gomez-Gil, "Leaf classification in sunflower crops by computer vision and neural networks," *Computers and Electronics in Agriculture*, vol. 78, no. 1, pp. 9–18, 2011.

[18] F. Ahmed, H. Al-Mamun, H. Bari, E. Hossain, and P. Kwan, "Classification of crops and weeds from digital images: a support vector machine approach," *Crop Protection*, vol. 40, pp. 98–104, 2012.

[19] J. Ambuel, T. S. Colvin, and D. L. Karlen, "A fuzzy logic yield simulator for prescription farming," *Transactions of the ASAE*, vol. 37, no. 6, pp. 1999–2009, 1994.

[20] S. Drummond, K. A. Sudduth, and S. J. Birrell, "Analysis and correlation methods for spatial data," in *ASAE Paper No. 95–1335. St. Joseph, Mich.: ASAE.*, 1995, pp. 350–359.

[21] S. Drummond, K. Sudduth, A. Joshi, S. Birrell, and N. Kitchen, "Statistical and neural methods for site-specific yield prediction," *Transactions of the ASAE*, vol. 46, no. 1, pp. 5–14, 2003.

[22] B. Ji, Y. Sun, S.Yang, and J. Wan, "Artificial neural networks for rice yield prediction in mountainous regions," *Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science.*, vol. 145, no. 3, pp. 249–261, 2007.

[23] G. RuB, M. S. R. Krus and, and P. Wagner, "Optimizing wheat yield prediction using different topologies of neural networks," in *Proceedings of IPMU-08*, 2008, pp. 576–582.

[24] G. RuB, "Data mining of agricultural yield data: A comparison of regression models," in *Industrial Conference on Data Mining ICDM 2009: Advances in Data Mining. Applications and Theoretical Aspects, P. Perner (Ed.), Lecture Notes in Artificial Intelligence 6171, Berlin, Heidelberg, Springer*, 2009, pp. 24–37.

[25] S. Fukuda, W. Spreer, E. Yasunaga, K. Yuge, V. Sardsud, and J. Muller, "Random forests modelling for the estimation of mango (mangifera indica l. cv.chok anan) fruit yields under different irrigation regimes," *Journal of Agricultural Water Management*, vol. 116, no. 3, pp. 142–150, 2013.

[26] I. Oliveira, R. Cunha, B. Silva, and M. Netto, "A scalable machine learning system for pre-season agriculture yield forecast," in *the 14th IEEE eScience Conference*, 2018.

[27] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, "Predictive ability of machine learning methods for massive crop yield prediction," *Spanish Journal of Agricultural Research*, vol. 12, no. 2, pp. 313–328, 2014.

[28] J. Jeong, J. Resop, N. Mueller, D. Fleisher, K.Yun, E. Butler, D. Timlin, K. Shim, J. Gerber, V. Reddy, and S. Kim, "Random forests for global and regional crop yield predictions," *PLoS ONE*, vol. 11, no. 6, 2016.

[29] L. Kouadio, R. Deo, V. Byrareddy, J. Adamowski, S. Mushtaq, and V. P. Nguyen, "Artificial intelligence approach for the prediction of robusta coffee yield using soil fertility properties," *Computers and Electronics in Agriculture*, vol. 155, pp. 324–338, 2018.

[30] P. Filippi, E. Jones, T. Bishop, N. Acharige, S. Dewage, L. Johnson, S. Ugbaje, T. Jephcott, S. Paterson, and B. Whelan, "A big data approach to predicting crop yield," in *Proceedings of the 7th Asian-Australasian Conference on Precision Agriculture 16–18 October 2017, Hamilton, New Zealand.*, 2017.

[31] X. Ye, K. Sakai, L. Garciano, S. Asada, and A. Sasao, "Estimation of citrus yield from airborne hyperspectral images using a neural network model," *Ecological Modelling*, vol. 198, no. 3-4, pp. 426–432, 2006.

[32] X. Pantazi, D. Moshou, T. Alexandridis, R. Whetton, and A. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," *Journal of Computers and Electronics in Agriculture*, vol. 121, pp. 57–65, 2016.

[33] R.Ji, J.Min, Y.Wang, H.Cheng, H.Zhang, and W.Shi, "In-season yield prediction of cabbage with a hand-held active canopy sensor," *Sensors*, vol. 17, no. 10, 2017.

[34] H. Cheng, L. Damerow, Y. Sun, and M. Blanke, "Early yield prediction using image analysis of apple fruit and tree canopy features with neural networks," *Journal of imaging*, vol. 3, no. 1, 2017.

[35] A. Wang, C. Tran, N. Desai, D. Lobell, and S. Ermon, "Deep transfer learning for crop yield prediction with remote sensing data," in *Proceedings of the COMPASS'18, Proceedings of the 1st ACM SIGCAS conference on Computing and Sustainable Societies. Menlo Park and San Jose, CA, USA. June 20-22*, 2018.

[36] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep gaussian process for crop yield prediction based on remote sensing data," in *the Thirty-First AAAI Conference on Artificial Intelligence. AAAI Publications*, 2017, pp. 4559–4566.

[37] K. Kuwata and R. Shibasaki, "Estimating crop yields with deep learning and remotely sensed data," in *Geoscience and Remote Sensing Symposium, IEEE International*, 2015, pp. 858–861.

[38] S. Mohanty, D. Hughes, and M. Salathe, "Using deep learning for image-based plant disease detection," *Technical Advances in Plant Science, a section of the journal Frontiers in Plant Science.*, vol. 7, pp. 1–10, 2016.

[39] D. Al-Bashish, M. Braik, and S.Bani-Ahmed, "Detection and classification of leaf disease using k-means-based segmentation and neural networks-based classification," *Information Technology. Asian Network of scientific information*, vol. 10, no. 2, pp. 267–275, 2011.

[40] Q. Yao, Z. Guan, Y. Zhou, J. Tang, Y. Hu, and B. Yang, "Application of support vector machine for detecting rice diseases using shape and color texture features," in *International Conference on Engineering Computation, IEEE computer society. 2-3 May 2009 Hong Kong, China*, 2009, pp. 79–83.

[41] K. Huang, "Application of artificial neural network for detecting phalaenopsis seedling diseases using color and texture features," *Computers and Electronics in Agriculture*, vol. 57, no. 1, pp. 3–11, 2007.

[42] Y. Tian, T. Li, C. Li, Z. Piao, G. Sun, and B. Wang, "Method for recognition of grape disease based on support vector machine," *Transaction. CSAE*, vol. 23, no. 6, pp. 175–180, 2007.

[43] S. Sladojevic, M. Arsenovic, A. A. D. Culibrk, and D. Stefanovic, "Deep neural networks based recognition of plant diseases by leaf image classification," *Computational Intelligence and Neuroscience*, vol. 2016, 2016.

[44] Z. Liu, H. Wu, and J. Huang, "Application of neural networks to discriminate fungal infection levels in rice panicles using hyperspectral reflectance and principal components analysis," *Computers and Electronics in Agriculture*, vol. 72, no. 2, pp. 99–106, 2010.

[45] M. El-Telbany and M. Warda, "An empirical comparison of tree-based learning algorithms: An egyptian rice diseases classification case study," *International Journal of Advanced Research in Artificial Intelligence*, vol. 5, no. 1, 2016.

[46] K. Ferentinos, "Deep learning models for plant disease detection and diagnosis," *Computers and Electronics in Agriculture*, vol. 145, pp. 311–318, 2018.

[47] B. Liu, Y. Zhang, D. He, and Y. Li, "Identification of apple leaf diseases based on deep convolutional neural networks," *Symmetry*, vol. 10, no. 1, 2018.

[48] K. Yamamoto, T. Togami, and N. Yamaguch., "Super-resolution of plant disease images for the acceleration of image-based phenotyping and vigor diagnosis in agriculture," *Sensors*, vol. 17, no. 11, 2017.

[49] E. Too, L. Yujian, S. Njuki, and L. Yingchun, "A comparative study of fine-tuning deep learning models for plant disease identification," *Computers and Electronics in Agriculture. In press*, 2018.

[50] A. Cruz, A. Luvisi, L. D. Bellis, and Y. Ampatzidis, "X-fido: An effective application for detecting olive quick decline syndrome with deep learning and data fusion," *Frontiers Plant Science*, vol. 8, 2017.

[51] S. Akbarzadeh, A. Paap, S. Ahderom, B. Apopei, and K. Alameh, "Plant discrimination by support vector machine classifier based on spectral reflectance," *Computers and Electronics in Agriculture*, vol. 148, pp. 250–258, 2018.

[52] J. Gao, D. Nuyttens, P. Lootens, Y. He, and J. Pieters, "Recognising weeds in a maize crop using a random forest machine-learning algorithm and near-infrared snapshot mosaic hyperspectral imagery," *Biosystems Engineering*, vol. 170, pp. 30–50, 2018.

[53] H. Habaragamuwa, Y. Ogawa, T. Suzuki, T. Masanori, and O. Kondo, "Detecting greenhouse strawberries (mature and immature), using deep convolutional neural network," *Engineering in Agriculture, Environment and Food*, vol. 11, no. 3, pp. 127–138, 2018.

[54] P. Ramos, F. Prieto, E. Montoya, and C. Oliveros, "Automatic fruit count on coffee branches using computer vision," *Computers and Electronics in Agriculture*, vol. 137, pp. 9–22, 2017.

[55] S. Amatya, M. Karkee, A. Gongal, Q. Zhang, and M. Whiting, "Detection of cherry tree branches with full foliage in planar architecture for automated sweet-cherry harvesting," *Biosystems Engeneering*, vol. 146, pp. 3–15, 2015.

[56] J. Senthilnath, A. Dokania, M. Kandukuri, K. Ramesh, G. Anand, and S. Omkar, "Detection of tomatoes using spectral-spatial methods in remotely sensed rgb images captured by uav," *Biosystems Engineering*, vol. 146, pp. 16–32, 2016.

[57] S. Sengupta and W. Lee, "Identification and determination of the number of immature green citrus fruit in a canopy under different ambient light conditions," *Biosystems Engineering*, vol. 117, pp. 51–61, 2014.

[58] I. Sa, Z. Ge, F. D. B. Upcroft, T. Perez, and C. Mccool, "Deepfruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, 2016.

[59] F. Kurtulmus, W. Lee, and A.Vardar, "Immature peach detection in colour images acquired in natural illumination conditions using statistical classifiers and neural network," *Precision Agriculture*, vol. 15, no. 1, pp. 57–79, 2014.

[60] S. Lee and L. Kerschberg, "Methodology and life cycle model for data mining and knowledge discovery in precision agriculture," in *the IEEE International Conference on Systems, Man and Cybernetics, Vol. 3*, 1998, pp. 2882–2887.

[61] E. Papageorgiou, A. Markinos, and T. Gemptos, "Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application," *Applied Soft Computing*, vol. 11, no. 4, pp. 3643–3657, 2011.

[62] ——, "Application of fuzzy cognitive maps for cotton yield management in precision farming," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12 399–12 413, 2009.

[63] V. Leemans and M. Destain, "A real-time grading method of apples based on features extracted from defects," *Journal of Food Engineering*, vol. 61, no. 1, pp. 83–89, 2004.

[64] G. Meyer, J. Neto, D. Jones, and T. Hindman, "Intensified fuzzy clusters for classifying plant, soil, and residue regions of interest from color images," *Computers and Electronics in Agriculture*, vol. 42, no. 3, pp. 161–180, 2004.

[65] J. Galambosova, V. Rataj, R. Prokeinova, and J. Presinska, "Determining the management zones with hierarchic and non-hierarchic clustering methods," *Special Issue in Research in Agriculture Engineering*, vol. 60, pp. 44–51, 2014.

[66] M. Ingeli, J. Galambosova, R. Prokeinova, and V. Rataj, "Application of clustering method to determine production zones of field," *Acta Technologica Agriculturae*, vol. 18, no. 2, pp. 42–45, 2015.

[67] E. Speranza, R. Ciferri, C. Grego, and L. Vicente, "A cluster-based approach to support the delination of management zones in precision agriculture," in *IEEE 10 th International Conference on eScience*, 2014.

[68] J. Martinez-Casasnovas, A. Escola, and J. Arno, "Use of farmer knowledge in the delineation of potential management zones in precision agriculture: A case study in maize (zea mays l.)," *Agriculture*, vol. 84, no. 8, 2018.

[69] A. Tagarakis, V. Liakos, S. Fountas, S. Koundouras, and T. Gemtos, "Management zones delineation using fuzzy clustering techniques in grapevines," *Precision Agriculture*, vol. 14, no. 1, pp. 18–39, 2013.

[70] L. Vendrusculo and A. Kaleita, "Modeling zone management in precision agriculture through fuzzy c-means technique at spatial database." in *Proceedings of the 2011 ASABE Annual International Meeting Sponsored by ASABE. Gault House, Louisville, Kentucky. August 7-10*, 2016, pp. 350–359.

[71] J. Ping, C. Green, K. Bronson, R. Zartman, and A. Dobermann, "Delineating potential management zones for cotton based on yields and soil properties," *Soil Science*, vol. 170, no. 5, pp. 371–385, 2005.

[72] X. Zhang, L. Shi, X. Jia, G. Seielstad, and C. Helgason, "Zone mapping application for precision farming: a decision support tool for variable rate application," *Precision Agriculture*, vol. 11, no. 2, pp. 103–114, 2010.

[73] A. Brock, S. Brouder, G. Blumhoff, and B. Hofmann, "Defining yield-based management zones for corn-soybean rotations," *Agronomy Journal*, vol. 97, no. 4, pp. 1115–1128, 2005.

[74] L. Yan, S. Zhou, W. Cifang, L. Hongyi, and L. Feng, "Classification of management zones for precision farming in saline soil based on multi-data sources to characterize spatial variability of soil properties," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 23, no. 8, pp. 84–89, 2007.

[75] B. Feldman, E. Martin, and T. Skotnes, "Big data in healthcare hype and hope, october 2012.dr. bonnie 360, 2012." 2012.

Performance Evaluation of Exteneded Iterative Identification Algorithm Based on Decomposed (CARARMA) Systems

Nasar Aldian Ambark Shashoa
Department of Electrical and Computer Engineering
Libyan Academy
Tripoli, Libya

Omar Abusaeeda
Department of computer engineering
Azzaytuna University
Tarhuna, Libya

Salah Mohamed Naas
Department of computer engineering
Azzaytuna University
Tarhuna, Libya

Ibrahim N. Jleta
Department of Electrical and Computer Engineering
Libyan Academy
Tripoli, Libya

***Abstract*— a parameter estimation algorithm for single-input single-output system, represented by (CARARMA) model is derived based on extended iterative identification algorithm. Two identification models is obtained, the first one is the system parameters model and the next is the noise parameters model, the purpose is to enhance the computational efficiencies of the model. The values of some statistical indicators, which are (r), (MBE), and (d) are determined to evaluate the model performance. In addition, Rissanen's minimum description length (MDL) method is used for a selection of system model order. The effectiveness of the algorithm is confirmed in the Simulation results.**

Keywords— parameter estimation, identification algorithm, Descriptive Statistics, statistical indicators, correlation coefficient

I. INTRODUCTION

In the past few decades, Parameter estimation or system identification has received extensive attention in control systems, chemical processes and signal processing, [1]. The identification/estimation of parameters of the system from an input–output sequences is called system identification or parameter estimation [2]. Many developed identification methods in the literature have been using for identification of the parameters of linear systems and as well as nonlinear systems [3], for instance, least-squares methods the maximum likelihood methods, and the iterative identification methods [1]. In general, most of the realistic physical processes are multivariable systems and then, many estimation methods for multivariable systems has been developed lately [4]. For example, a R algorithm for the multi-input single-output systems based on the bias compensation technique was presented by Zhang; decomposition based maximum likelihood generalized extended least squares algorithm for multiple-input single-output nonlinear Box–Jenkins systems was derived by Chen and Ding. Multivariable controlled autoregressive moving average (ARMA) systems identification have Difficulties, because the information vector contains unknown variables and unmeasurable noise terms. The usage of iterative identification is the solution for this problem and these unmeasurable variables are replaced with their estimates, and a least squares based iterative algorithm is proposed in, for multivariable controlled ARMA systems. [5]. In these systems, a hierarchical identification principle and iterative identification principle is combined and the aim is to decompose the original identification problem into two sub issues with smaller sizes of the covariance matrices is used [6]. First sub issues is system identification model and the next is identification model of the noise in order to enhance computational efficiencies [5]. The performance of models need to test and there are many statistical indicators are commonly used for that, such as correlation coefficient, Mean Bias Error (MBE), and d statistics [7]. This paper is organized as follows:. Section 2, identification model of the proposed algorithm is derived. Section 3, model validation performance evaluation is discussed. In Section 4, simulation results is presented. Finally, concluding remarks are offered in section 5.

II. THE IDENTIFICATION MODEL DERIVATION OF THE PROBOSED ALGORITHM

This work considers single-input single-output system, represented by (CARARMA) model, as shown in fig. 1

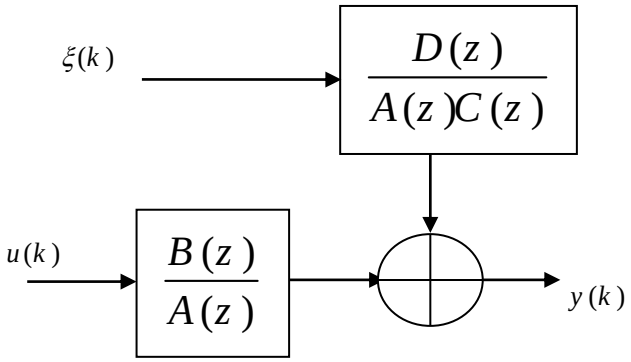


Fig. 1. Structure of the proposed Model

$$A(z^{-1})y(k) = B(z^{-1})u(k) + \frac{D(z^{-1})}{C(z^{-1})}\xi(k) \quad (1)$$

$$v(k) = \frac{D(z^{-1})}{C(z^{-1})}\xi(k) \quad (2)$$

Where

$$A(z^{-1}) = 1 + a_1 z^{-1} + a_2 z^{-2} + ... + a_n z^{-n}$$
$$B(z^{-1}) = b_1 z^{-1} + b_2 z^{-2} + ... + b_n z^{-n}$$
$$C(z^{-1}) = 1 + c_1 z^{-1} + c_2 z^{-2} + ... + c_n z^{-n}$$
$$D(z^{-1}) = 1 + d_1 z^{-1} + d_2 z^{-2} + ... + d_n z^{-n}$$

are the polynomials in the unit backward shift operator as

$z^{-1}y(k) = y(k-1)$ [8]. $u(k)$, $y(k)$ and $\xi(i)$ are the input, output and noise while a_i, b_i and d_i are the parameters. This algorithm can be written as

$$y(k) = -\sum_{i=1}^{n} a_i y(k-i) + \sum_{i=1}^{n} b_i u(k-i) + v(k) \quad (3)$$

Decomposed technique is derived to obtain: first, the System identification model is

$$y(k) = -\sum_{i=1}^{n} a_i y(k-i) + \sum_{i=1}^{n} b_i u(k-i) \quad (4)$$

The system parameter is defined as $\theta_a = [a_1, ...a_n]^T$ $\theta_b = [b_1, ..., b_n]^T$

$$\theta_s = [\theta_a \quad \theta_b]^T$$

The system information vectors are [7].

$Z_a{}^T = [-y(i-1)...-y(i-n)]^T$. $Z_b{}^T = [u(i-1)..u(i-n)]$.

$$Z_s^T(i) = [Z_a^T \quad Z_b^T] \quad (5)$$

Then, the system output is

$$y_s(k) = Z_s^T(k)\theta_s \quad (6)$$

Next, the noise identification model is

$$v(k) = -\sum_{i=1}^{n} c_i v(k-i) + \sum_{i=1}^{n} d_i \xi(k-i) + \xi(k) \quad (7)$$

Define the noise parameter vectors and the information vectors as

$\theta_c = [c_1, c_2..., c_n]^T$, $\theta_d = [d_1, d_2..., d_n]^T$

$\theta_n = [\theta_c \quad \theta_d]^T$

$Z_c^T(k) = [-v(k-1)...-v(k-n)]$, $Z_d^T(k) = [\xi(k-1)..\xi(k-n)]$

$$Z_n^T(i) = [Z_c^T \quad Z_d^T] \quad (8)$$

The linear regression form is

$$v(k) = Z_n^T(k)\theta_n + \xi(k) \quad (9)$$

This equation and equation (6) can be substituting into equation (3) and the result gives the following identification algorithm [9].

$$y(k) = [Z_s^T \quad Z_n^T]\begin{bmatrix}\theta_s \\ \theta_n\end{bmatrix} + \xi(k)$$

$$y(k) = Z_s^T\theta_s + Z_n^T(k)\theta_n + \xi(k) = Z^T\theta + \xi(k) \quad (10)$$

Where

$$Z^T = \begin{bmatrix}Z_s^T & Z_n^T\end{bmatrix} \quad \text{and} \quad \theta = \begin{bmatrix}\theta_s \\ \theta_n\end{bmatrix}$$

Define two intermediate variables:

$$y_1(k) = y(k) - Z_n^T\theta_n \quad (11)$$

$$y_2(k) = y(k) - Z_s^T\theta_s \quad (12)$$

From equation (10), we can written equation (11) and equation (12) as

$$y_1(k) = Z_s^T\theta_s + \xi(k) \quad (13)$$

$$y_2(k) = Z_n^T\theta_n + \xi(k) \quad (14)$$

the data from $k = 1$ to $k = N$ is considered. The output vectors $Y(N), Y_1(N)$ *and* $Y_2(N)$, the regression vectors $\Xi_s(N)$, $\Xi_n(N)$ and noise vector $\Im(N)$ can be written as

$$Y(N)=\begin{bmatrix} y(1) \\ . \\ . \\ . \\ y(N) \end{bmatrix},\ Y_1(N)=\begin{bmatrix} y_1(1) \\ . \\ . \\ . \\ y_1(N) \end{bmatrix},\ Y_2(N)=\begin{bmatrix} y_2(1) \\ . \\ . \\ . \\ y_2(N) \end{bmatrix}$$

$$\Xi_s(N)=\begin{bmatrix} Z_s(1) \\ . \\ . \\ . \\ Z_s(N) \end{bmatrix},\ \Xi_n(N)=\begin{bmatrix} Z_n(1) \\ . \\ . \\ . \\ Z_n(N) \end{bmatrix},\ \Im(N)=\begin{bmatrix} \xi(1) \\ . \\ . \\ . \\ \xi(N) \end{bmatrix}$$

From (11) and (12), we have [10].

$$Y_1(N)=Y(N)-\Xi_n^T\theta_n \tag{15}$$

$$Y_2(N)=Y(N)-\Xi_s^T\theta_s \tag{16}$$

and from (13) and (14)

$$Y_1(N)=\Xi_s^T\theta_s+\Im(N) \tag{17}$$

$$Y_2(N)=\Xi_n^T\theta_n+\Im(N) \tag{18}$$

The equation error (residual) $\xi(k)$ is introduced as

$$\xi_1(k)=y_1(k)-Z_s{}^T(k)\theta_s \tag{19}$$

$$\xi_2(k)=y_2(k)-Z_n{}^T(k)\theta_n \tag{20}$$

Equations (19) and (20) can be written compactly as

$$\Im_1=Y_1-\Xi_s^T\theta_s \tag{21}$$

$$\Im_2=Y_2-\Xi_n^T\theta_n \tag{22}$$

For the minimization of the error vector $\xi(k)$, the least-square method can applied. To this end, define the following cost function [11].

$$J_1=\Im_1^T\Im_1=\sum_{i=1}^{N}\xi_1^2(k) \tag{23}$$

$$J_2=\Im_2^T\Im_2=\sum_{i=1}^{N}\xi_2^2(k) \tag{24}$$

If equation (21) is substituted in equation (23) and equation (22) is substituted in equation (24), we obtain

$$J_1=(Y_1-\Xi_s^T\theta_s)^T(Y_1-\Xi_s^T\theta_s) \tag{25}$$

$$J_2=(Y_2-\Xi_n^T\theta_n)^T(Y_2-\Xi_n^T\theta_n) \tag{26}$$

Hence

$$\frac{\partial J_1}{\partial\theta_s}=-2Z_s^T(Y_1-\Xi_s\theta_s) \tag{27}$$

$$\frac{\partial J_2}{\partial\theta_n}=-2Z_n^T(Y_2-\Xi_n\theta_n) \tag{28}$$

Where the following formula used

$$\frac{\partial}{\partial\Theta}[A\Theta]=\frac{\partial}{\partial\Theta}[\Theta^TA^T]=A^T \tag{29}$$

If we set $\partial J/\partial\theta$ equal zero, we obtain

$$\Xi_s^TY_1=\Xi_s^T\Xi_s\theta_s \tag{30}$$

$$\Xi_n^TY_2=\Xi_n^T\Xi_n\theta_n \tag{31}$$

Relations (23) and (24) have solutions when the matrixes $\Xi_s^T\Xi_s$ and $\Xi_n^T\Xi_n$ are invertible, and therefore, the results are

$$\hat{\theta}_s=[\Xi_s^T\Xi_s]^{-1}\Xi_s^TY_1 \tag{32}$$

$$\hat{\theta}_n=[\Xi_n^T\Xi_n]^{-1}\Xi_n^TY_2 \tag{33}$$

Equations (15) and (16) can be substituted into equations (32) and (33) respectively as

$$\hat{\theta}_s=[\Xi_s^T\Xi_s]^{-1}\Xi_s^T\ [Y(N)-\Xi_n^T\theta_n] \tag{34}$$

$$\hat{\theta}_n=[\Xi_n^T\Xi_n]^{-1}\Xi_n^T\ [Y\ (N)-\Xi_s^T\theta_s] \tag{35}$$

Because $Z_n(k)$ in $\Xi_n(k)$ contains the unmeasurable noise terms $v(k)$ and $\xi(k)$, it is imp0ssible to calculate $\hat{\theta}_s$ and $\hat{\theta}_n$. Here, the interactive estimation theory is the solution to generate θ_s, θ_n. $k = 1,2,...$ be an iteration variable $\hat{\theta}_{sk}$ and $\hat{\theta}_{nk}$ be estimates of θ_s and θ_n, $\hat{v}(k-1)$ and $\hat{\xi}(k-1)$ be the estimates of $v(k-1)$ and $\xi(k-1)$, and $\hat{Z}_n(k)$ at iteration k obtained by replacing $v(k-1)$ and $\xi(k-1)$ in $Z_n(k)$ with $\hat{v}(k-1)$ and $\hat{\xi}(k-1)$ [12], i.e.

$$\hat{Z}_{nk}^T(k) = [-\hat{v}_{k-1}(k-1)...-\hat{v}_{k-1}(k-m), \hat{\xi}_{k-1}(k-1)...\hat{\xi}_{k-1}(k-m)] \quad (36)$$

$$\hat{\Xi}_{nk}(N) = \begin{bmatrix} \hat{Z}_{nk}(1) \\ . \\ . \\ . \\ Z_{nk}(N) \end{bmatrix}$$

From equation (18), we have

$$\xi(k) = y(k) - Z_s^T\theta_s - Z_n^T(k)\theta_n \quad (37)$$

$Z_n(k)$, θ_s and θ_n replaced with, $\hat{Z}_{nk}^T(k)$ $\hat{\theta}_{sk}$ and $\hat{\theta}_{nk}$, and therefore $\bar{\xi}(k)$ of $\xi(k)$ is calculated by

$$\hat{\xi}(k) = y(k) - Z_s^T\hat{\theta}_{sk} - \hat{Z}_{nk}^T(k)\hat{\theta}_{nk} \quad (38)$$

$\Xi_n(N)$, θ_s and θ_n are replaced with $\hat{\Xi}_n(N)$, $\hat{\theta}_{sk}$ and $\hat{\theta}_{nk}$. The estimated θ_s and θ_n based on two-stage (CARMA) based RLS algorithm for SISO systems as

$$\hat{\theta}_s = [\Xi_s^T\Xi_s]^{-1}\Xi_s^T\ [Y(N) - \hat{\Xi}_{nk}^T\hat{\theta}_{nk}] \quad (39)$$

$$\hat{\theta}_n = [\hat{\Xi}_{nk}^T\hat{\Xi}_{nk}]^{-1}\hat{\Xi}_{nk}^T\ [Y(k) - \Xi_s^T\hat{\theta}_{sk}] \quad (40)$$

$$Z_s^T(k) = [-y(k-1)...-y(k-n), u(k-1)..u(k-m)]$$

$$\hat{Z}_{nk}^T(k) = [-\hat{v}_{k-1}(k-1)...-\hat{v}_{k-1}(k-m), \hat{\xi}_{k-1}(k-1)..\hat{\xi}_{k-1}(k-m)]$$

From equation (18)

$$\hat{\xi}(k) = y(k) - Z_s^T\hat{\theta}_{sk} - \hat{Z}_{nk}^T\hat{\theta}_{nk} = y(t) - [Z_s^T \quad \hat{Z}_{nk}^T]\begin{bmatrix} \hat{\theta}_{sk} \\ \hat{\theta}_{nk} \end{bmatrix}$$

$$\hat{\xi}(k) = y(k) - \hat{Z}^T\hat{\theta}_k \quad (41)$$

III. MODEL PERFORMANCE AND ORDER SELECTION

Many Statistical Indicators have been used as a standard statistical metrics to measure model performance such as correlation coefficient, the mean bias error and Descriptive Statistics (d). In addition to, there are many methods in the literature are using the selection of model order. One of these methods is Rissanen's minimum description length (MDL).

A. Correlation Coefficient

Correlation is a statistical method that measures the degree of connection between quantitative variables. This degree of relationship can vary from none weak or strong. A correlation coefficient is the a correlation analysis result and is given by

$$r = \frac{\sum_{k=1}^{N}(\hat{y}(k) - \bar{\hat{y}}(k))(y(k) - \bar{y}(k))}{\sqrt{\sum_{k=1}^{N}(\hat{y}(k) - \bar{\hat{y}}(k))^2 \sum_{k=1}^{N}(y(k) - \bar{y}(k))^2}} \quad (42)$$

Where,

$\hat{y}(k)$ = estimated data

$\bar{\hat{y}}(k)$ = average value of estimated data

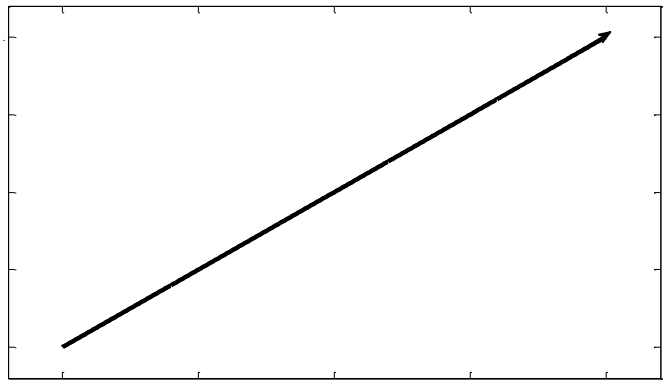
$y(k)$ = measured data

$\bar{y}(k)$ =average value of measured data

N =number of the sequences

The values range of a correlation coefficient is always between -1 and +1. If it is +1, that means perfect linear relationship in linear positive between two variables. If it is -1, perfect linear relationship in linear negative between two variables and if it is 0, that means there is no linear relationship between the variables as illustrate in Fig.2 [13].

Positive Correlation



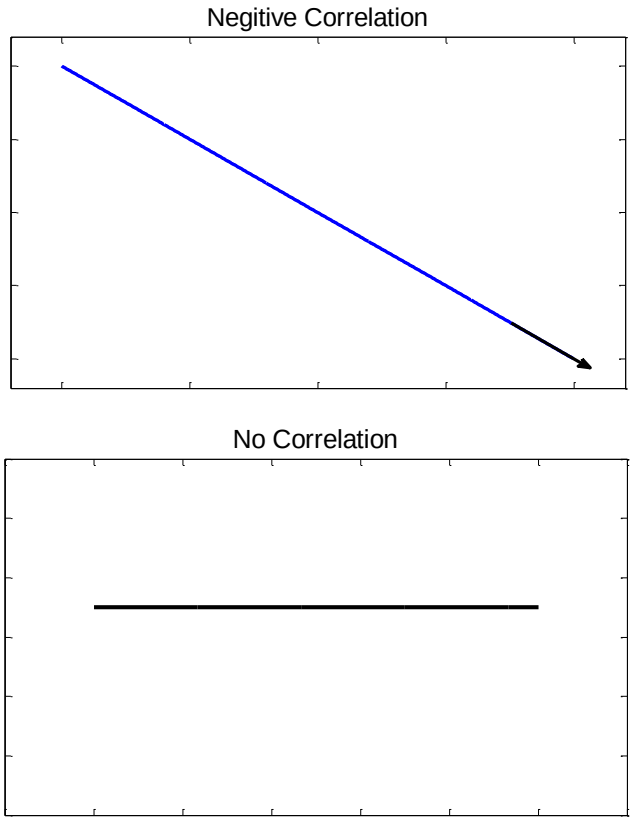


Fig. 2. Correlation Between Two Variables

B. *Mean Bias Error (MBE)*

MBE is the difference between the mean of the estimated and true data and it computed by

$$MBE = \frac{1}{N}\sum_{k=1}^{N}(\hat{y}(k)-y(k)) \tag{43}$$

Low value MBE is preferred [7].

C. *Descriptive Statistics (d)*

A descriptive statistics "d" measure of the degree to which the model predictions are error free. 'd' values are varies between zero and one. If the computed value is (1), this point out the perfect agreement between the estimated and true data whereas if the computed value is (0), this point out complete disagreement. A descriptive statistics is computed by

$$d = 1-\frac{\sum_{k=1}^{N}(\hat{y}(k)-y(k))^2}{\sum_{k=1}^{N}[|\hat{y}(k)-\bar{y}(k)|+|y(k)-\bar{y}(k)|]^2} \tag{44}$$

D. *Rissanen's minimum description length (MDL)*

(MDL) is one of the methods that used for validation of selection of the model order and is represented by

$$MDL = (1+\log N * n/N)V \tag{45}$$

Where

N is the samples number, n is the model parameters number and V is the variance of model residuals [14].

IV. Simulation Results

In order to testing the proposed model performance using the statistical indicators that explained in section 3, consider the following example as a second order system

$$A(z^{-1}) = 1+0.28z^{-1}+0.95z^{-2} \qquad B(z^{-1}) = -0.93z^{-1}+0.68z^{-2}$$

$$C(z^{-1}) = 1+0.33z^{-1} \qquad D(z^{-1}) = 1-0.48z^{-1}$$

$$\hat{\theta}_s(k) = [0.28, 0.95, -0.93, 0.68]^T \qquad \hat{\theta}_n(k) = [0.33, -0.48]^T$$

$u(k)$ is the input with $m = 0$ and $\sigma^2 = 1$, whilst $\xi(k)$ is the noise with $m = 0$ and $\sigma^2 = 0.4$. First, correlation coefficient is calculated for this model and data sequences from n =0 to n =1000 has been used. Correlation coefficient (r) for this model equal 99.93%. Fig. 2 shows the estimated output and the true output sorted in ascending order. The figure illustrates that they are almost perfectly related in linear positive.

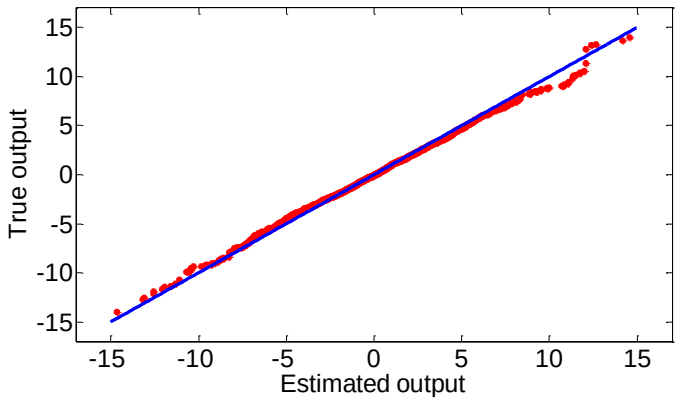


Fig. 3. the Estimated Output and the True Output

Next, Mean Bias Error is computed for the model and the result that has been obtained equal (0.0163). The result of this indicator shows that the model is good. Finally, Descriptive Statistics (d) indicator is determined for the proposed model and the value of this indicator equal (0.9983). This value indicates that the proposed model is nearly ideal. *MDL* is computed for various candidate models (from one to seven). The next figure shows Rissanen's minimum description length versus model order.

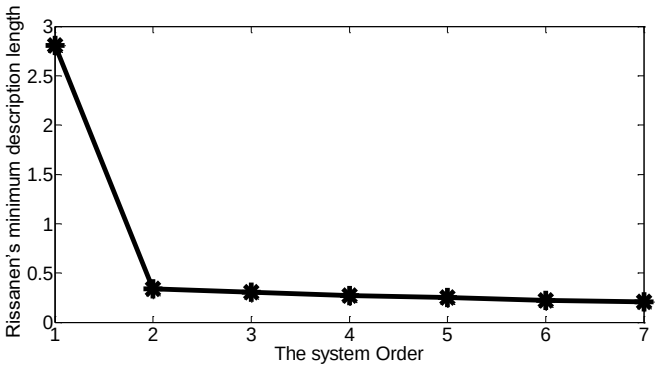


Fig. 4. Rissanen's minimum description length versus model order

The figure shows that the best model order is the second, and it is the same model order that we assumed. That mean, the validation of model order is high.

V. CONCLUSIONS

Modified iterative identification algorithm based on the decomposed technique for (CARARMA) systems has derived in this paper. The decomposition technique is used for estimating the parameters of the system and the noise parameters and therefore, the computational efficacy has been improved. In the simulation results, some statistical indicators have been calculated to evaluate the model performance and their values indicate that this model has high capability. In addition to, the validation of the model order has done using MDL method and the result indicates that the model has high validation of model order. The algorithm can be extended to MISO systems.

REFERENCES

[1] C. Wang, T. Tang, and D. Chen, "Least-Squares Based and Gradient Based Iterative Parameter Estimation Algorithms for a Class of Linear-in-Parameters Multiple-Input Single-Output Output Error Systems," Journal of Applied Mathematics. vol. 2014, 2014.

[2] L. Chen, J. Li and R. Ding, "Identification for the second-order systems based on the step response," Mathematical and Computer Modelling ,vol. 53, pp. 1074–1083, 2011.

[3] W. Xiong, W. Fan, and R. Ding, "Least-Squares Parameter Estimation Algorithm for a Class of Input Nonlinear Systems," Journal of Applied Mathematics, vol. 2012, 2012.

[4] F. Chen, F. Ding, "The filtering based maximum likelihood recursive least squares estimation for multiple-input single-output systems," Applied Mathematical Modelling., 2015.

[5] S.Sundari and A. Nachiappan, "Online Identification Using RLS Algorithm and Kaczmarz's Projection Algorithm for a Bioreactor Process ," International Journal of Engineering and Computer Science, vol. 3, pp. 7974-7978, September 2014.

[6] F. Chen,F. Ding, "Recursive Least Squares Identification Algorithms for Multiple-Input Nonlinear Box–Jenkins Systems Using the Maximum Likelihood Principle," Journal of Computational and Nonlinear Dynamics, vol. 11 / 021005-1, 2016.

[7] E.O. Falayi, J.O. Adepitan and A.B. Rabiu, "Empirical Models for the Correlation of Global Solar Radiation with Meteorological Data for Iseyin, Nigeria," The Pacific Journal of Science and Technology, vol. 9. No. 2. November 2008.

[8] W. Huang and F. Ding, " Coupled Least Squares Identification Algorithms for Multivariate Output-Error Systems,"Algorithms, January 2017.

[9] Y. Wang ,L. Xu and F. Ding, "Data filtering based parameter estimation algorithms for multivariable Box-Jenkins-like systems geria," 9th International Symposium on Advanced Control of Chemical Processes, June 7-10, 2015, Whistler, British Columbia, Canada.

[10] F.Ding, "Two-stage least squares based iterative estimation algorithm for CARARMA system modeling," Applied Mathematical Modelling, 2012.

[11] P.N., Paraskevopoulos Modern Control Engineering, NewYork, Marcel Dekker, 2002.

[12] Y. Liu, D.Wang and F. Ding, "Least squares based iterative algorithms for identifying Box–Jenkinsmodels with finite measurement data," Digital Signal Processing, vol. 20, pp. 1458–1467, 2010.

[13] N.J. Gogtay and U.M. Thatte., "Principles of Correlation Analysis," Journal of The Association of Physicians of India, vol. 65, March 2017.

[14] R.G.K.M. Aarts, "System Identification And Parameter Estimation", Course Edition: 2011/2012.

Definition Framework of Educational Process Construction Supported by an Intelligent Tutoring System

Walid Bayounes
RIADI Research Laboratory
ENSI Manouba University
Manouba, Tunisia
oualid.bayounes@iseahz.u-tunis.tn

Inès Bayoudh Sâadi
RIADI Research Laboratory
ENSI Manouba University
Manouba, Tunisia
ines.bayoudh@ensi.rnu.tn

Hénda Ben Ghézala
RIADI Research Laboratory
ENSI Manouba University
Manouba, Tunisia
henda.benghezala@riadi.rnu.tn

***Abstract*— The motivation behind this paper is the need for educational processes that can be constructed and adapted to the needs of the learners, the preferences of the tutors, the requirements of the system administrators and the system designers of ITS. Within this context, the paper explores the theory to study the problem of educational process construction. This study introduces a new multi-level view of educational processes. Based on the proposed view, a faceted definition framework conducts a comparative study between different ITS to understand and classify issues in educational process construction.**

Keywords—Educational Processes, Process Engineering, Adaptive Learning, Intelligent Tutoring System

I. INTRODUCTION

One of the evolving areas that would certainly occupy computer scientists in the next decade is computer supported learning environment. The latter is increasingly mediated by new technologies of information and communication. In fact, it can provide various technological support to guide teaching and learning. For that, this context is considered as multidisciplinary, which includes computer science, cognitive psychology, pedagogy, didactics and educational sciences.

There are different types of environments, including the traditional e-learning system, instructional design system and intelligent tutoring system. The latter is increasingly gaining popularity in the academic community because of their several learning benefits. This work concerns more particularly the process adaptation in intelligent tutoring systems. In fact, most of these systems allow only the application of content adaptation and neglect the adaptation of educational processes (learning process and pedagogical process) [1]. This is a major constraint for providing personalized learning path and appropriate learning content and for exploiting the richness of individual differences of the learning needs and the pedagogical preferences [2] [3].

In fact , the main research problem is the construction of the learning process in ITS. The study of this problem conducts to the extension of construction in order to take into consideration the teaching process, also called pedagogical process, which is strongly correlated with it. In the rest of the paper, the term “educational processes” will be used to indicate both learning and pedagogical process.

Within this context, literature survey is conducted by using a framework of educational process construction. The comparison framework is used in many engineering works in literature and has proven its efficiency in improving the understanding of various engineering disciplines (method engineering, process engineering,) [4]. By adopting the multi-levels view of educational process definition, the goal of the proposed framework is to identify the suited construction approach supported by an ITS in order to satisfy the individual learning needs and to respect the system constraints.

This paper is organized as follows. A description of the new multi-level view of educational processes is presented in the next section. In section three, we specify the definition framework. Our comparison framework is applied on selected intelligent tutoring systems in the fourth section. The section five concludes this work with our contribution and research perspectives.

II. MULTI-LEVELS VIEW OF EDUCATIONAL PROCESS

The most important task of this research is to study the different definitions of educational processes. In fact, valid process definition ensures valid process modeling to support an appropriate process adaptation.

The analysis of this study conducts the specification of new multi-levels of definition view (see Fig. 1). These levels are separated into the Psycho-Pedagogical Level, the Didactic Level, the Situational Level and the Online Level. The two first levels define the educational process in the theoretical layer. The two last levels define the practical layer. Both layers consider the correlation between pedagogical and learning facets.

The different definition levels are specified by three major components (Affective, Cognitive and Metacognitive). The first component defines the affective objective achieved by the educational process (why ?). The second component defines the cognitive product adopted by the process (what ?). The third component the metacognitive process used to achieve the affective objective (how ?).

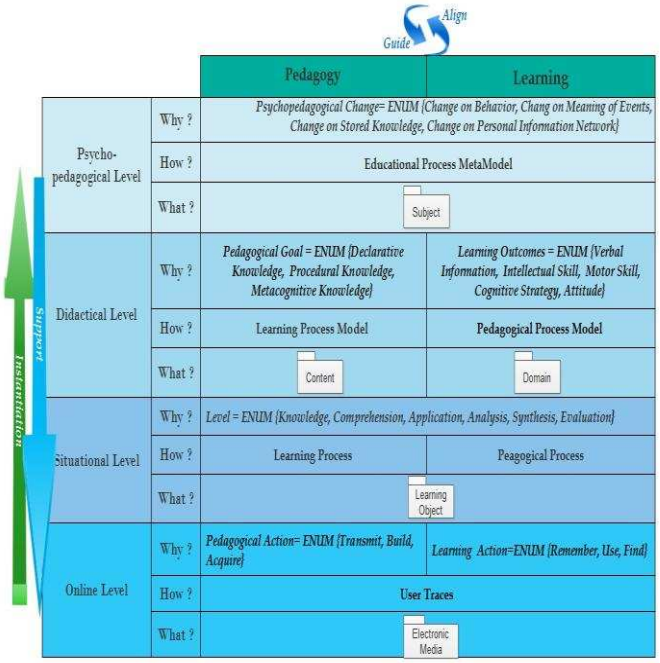


Fig. 1. Multi Levels View of Educational Processes

A. Definition Levels

1) Psycho-Pedagogical Level: At this level, the pedagogical and the learning view of process are not defined. The educational process is viewed as a psycho-pedagogical change that occurs by using the educational process meta model. The latter adopts various subjects, which are defined by different theories of learning paradigms.

2) Didactic Level: At this level, the process is specified by adopting the interaction between the pedagogical and the learning view. By considering the latter, the process model is used to achieve learning outcomes by considering the constraints of learning domain. For the pedagogical view, the process is viewed as pedagogical goal achieved by defining a process model based on pedagogical content.

3) Situational Level: It is the level of the instantiation of process models by considering the different characteristics of the learning/teaching situation in order to reach the desired learning level by using different learning objects. In fact, the situation characteristics are the objective, the different tasks and the different available resources.

4) Online Level: It is the level of the execution of different learning and pedagogical actions that are supported by different learning systems. This execution is achieved by adopting different electronic media in order to use the different learning objects.

B. Relationships

This multi-level view specifies the guidance/alignment relationships between learning and pedagogical facet and the instantiation/support relationships between different levels. In fact, they are used to support products and process relationships.

1) Products Relationships: The relationships between the cognitive products of the different levels are introduced (see Fig. 2). In fact, the subject is refined by the pedagogical content which specifies the didactic domain. This content is composed of different learning objects which are supported by various electronic media.

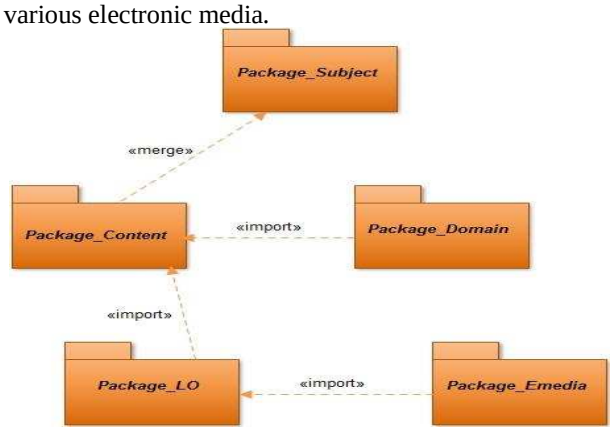


Fig. 2. Products Relationships

2) Processes Relationships: The two major processes relationships are guidance and alignment. On the one hand, the guid- ance relationship permits the monitoring of the learning process by the pedagogical process in order to satisfy the pedagogical preferences. For example, the tutor monitors the discussion between different learners about a case study project. On the other hand, the alignment relationships allow the orienting of the pedagogical process by the learning process in order to achieve the individual learning requirements. For instance, the result of pretest of an activist learner orients the tutor to start the explanation by using different study examples before presenting the definition of concept.

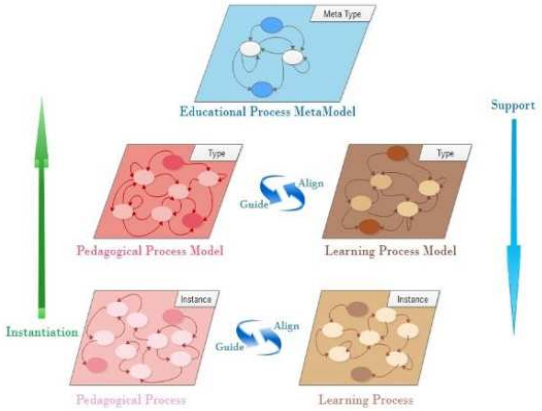


Fig. 3. Processes Relationships

III. PROPOSED FRAMEWORK

The issue of adaptation in ITS has become an important topic of research in recent years. With emergence of ITS, it has become possible to provide a learning process which

matches the learner characteristics, the pedagogical preferences and the specific learning needs. Within this context, literature survey is conducted by using a framework of educational process construction.

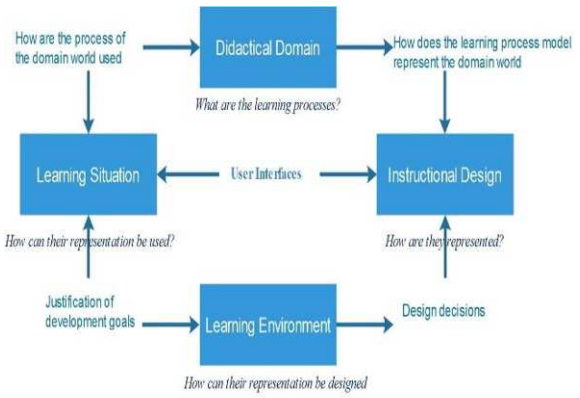


Fig. 4. Framework Worlds [5]

A. Didactic Domain World

It is the world of processes by considering the notion of process and its nature [5]. The didactic domain world contains the knowledge of domain about which the proposed process has to provide learning [5]. This world specifies process nature by considering the domain dimensions. Table I specifies a nature view by presenting three facets namely pedagogy, learning and process.

The pedagogy facet presents three attributes: the pedagogical orientation, the pedagogical method and the correlation. The different orientations are teacher-directed, learner-directed, and teacher-learner negotiated. The pedagogical method specifies the activities of learning process. It can be classified as direct instruction, indirect instruction, interactive instruction, experiential learning and independent study. The third attribute tests the correlation between pedagogical and learning process.

The learning facet presents three attributes: the performance, the learning mode and the learning outcome. The different performance types are remember, use and find [6]. These performances are accomplished by three different learning modes namely accretion, structuring and tuning [7].

TABLE I. NATURE VIEW

Facet	Definition	
	Attribute	***Values***
Pedagogy	Orientation	ENUM {Teacher-Directed, Learner-Directed, Teacher-Learner Negotiated}
	Method	ENUM {Not Defined, Direct, Indirect, Interactive, Independent Study, Experiential Learning}
	Correlation	ENUM {Considered, Not Considered}
Learning	Performance	ENUM {Remember, Use, Find}
	Mode	ENUM {Accretion, Structuring, Tuning}
	Outcome	ENUM {Verbal Information, Intellectual Skill, Motor Skill, Cognitive Strategy, Attitude}
Process	Type	ENUM {Strategic, Tactic, Implementation}
	Form	ENUM {Linear, Not Linear}

B. Instructional Design World

This world deals with the representation of processes by adopting predefined models. It focuses on the description view by respecting the notation conditions and the constraints of representation level. Table II presents the notation and the level of process description.

In the notation facet, four attributes are found : the type, the form, the content and the models. The two notation types are standard or proprietary. These types are used by adopting three major forms namely informal, semi-formal and formal. According to component display theory [6], these forms present four major types of content. These types are fact, concept, procedure and principle. These contents are used to instantiate the domain model, the learner model and the pedagogical model.

TABLE II. DESCRIPTION VIEW

Facet	Definition	
	Attribute	***Values***
Notation	Type	ENUM {Standard, Proprietary}
	Form	ENUM {Informal, Semi Formal, Formal}
	Content	ENUM {Fact, Concept, Procedure, Principle}
	Models	SET(ENUM {Domain Model, Learner Model, Pedagogical Model})
Level	Performance	ENUM {Course, Sequence, Activity}
	Mode	ENUM {Primitive, Generic, Aggregation}
	Outcome	ENUM {Activity, Product, Decision, Context}

Furthermore, the level facet presents three attributes: the granularity, the modularization and the coverage. The different levels of granularity are course, sequence of activities and simple activity. Three types of modularization namely primitive, generic and aggregation define these levels. In fact, the modularization attribute is used to capture the laws governing the learning process construction by using process modules [5]. This process has been defined differently in different coverage [2]:

- The activity: the process is defined as a related set of activities conducted for the specific purpose of product definition.
- The product: the process is a series of activities that cause successive product transformations to reach the desired product.
- The decision: the process is defined as a set of related decisions conducted for the specific purpose of product definition.
- The context: the process is a sequence of contexts causing successive product transformations under the influence of a decision taken in a context.

C. Learning Environment World

This world deals with the entities and activities, which arise as part of the engineering process itself [5]. It focuses on the design view by describing the implementation of process representation. Table III specifies a design view by presenting five facets namely context, construction, optimization, guidance and adaptation.

The first facet defines the type and the form of context. The construction facet deals with four issues: the approaches,

the methods, the tools and the techniques of adaptive learning process construction. In a manner analogous to the adaptation spectrum [8], one can organize construction approaches in a spectrum ranging from "low" flexibility to "high". The construction approach attribute is defined as Approach: ENUM {Rigid, Contingency, On-The-Fly}. These approaches are adopted by using the instantiation, the assembly and the ad hoc method. These methods apply three major techniques (curriculum sequencing, intelligent solution analysis, problem-solving support).

TABLE III. DESIGN VIEW

Facet	Definition	
	Attribute	***Values***
Context	Type	ENUM {Certain, Uncertain}
	Form	ENUM {Evolved, Stable}
Construction	Approach	ENUM {Rigid, On-The-Fly, Contingency}
	Method	ENUM {Instantiation, Assembly, Ad hoc}
	Technique	ENUM {Curriculum Sequencing, Intelligent Solution Analysis, Problem Solving Support}
Optimization	Model	ENUM {Not Defined, MAP-Based, Network-Based, Tree-Based}
	Method	ENUM {Exact, Heuristic, Meta heuristic}
	Technique	ENUM {Rule-Based, Case-Based}
	Parameters	SET(ENUM {Not Defined, Achieved Learning Intention, Achieved
Guidance	Type	ENUM {Strict, Flexible}
	Method	ENUM {Expository, Collaborative, Discovery}
	Parameters	SET(ENUM {Not Defined, Learning Style, Cognitive State, Teaching Style})
	Outcome	ENUM {Activity, Resource, Intention}
Adaptation	Dimension	SET(ENUM {Content, Structure, Presentation})
	Position	ENUM {User initiated adaptability, User desired adaptability supported by tools, User selection adaptation for system suggested features, System initiated adaptativity with pre-information to the user about the change, System initiated adaptivity}
	Method	ENUM {Macro adaptive, Aptitude-treatment interaction, Micro adaptive, Constructivistic-collaborative}
	Technique	ENUM {Adaptive Interaction, Adaptive Course Delivery, Content Discovery and Assembly, Adaptive Collaboration Support}
	Parameters	SET(ENUM {Learning Goal, Learning History, Prior Knowledge, System Information})

In the optimization facet, four attributes are found: the model, the method, the technique and the parameters of optimization. The different models are map-based, network-based and tree- based. Three types of methods namely exact, heuristic, and meta heuristic apply these models. These methods use rule-based and case-based techniques. The achieved educational intentions can be used to implement these optimization techniques.

In addition, the guidance facet defines four attributes: the type, the method, the outcomes and the parameters of guidance. The guidance types are strict and flexible. These types use three major methods namely expository, collaborative and discovery. The outcomes of these methods are activity, resource and intention. To this end, the guidance methods adopt three parameters learning style, cognitive state and teaching style.

Finally, the fourth facet introduces the adaptation by defining five attributes, which are the dimension, the position, the method, the technique and the parameters of adaptation. The three dimensions of adaptation are content, structure and presentation. The position is identified by adopting the adaptation spectrum [8]. Each position is satisfied by using four major methods namely macro adaptive, aptitude treatment interaction, micro adaptive, and constructivistic-collaborative [9]. These methods are implemented by four major techniques, which are adaptive interaction, adaptive course delivery, content discovery and assembly and adaptive collaboration support [10]. These techniques adopt four major parameters namely learning goal, learning history, prior knowledge and system information.

D. Learning Situation World

This world supports the scenario view by examining the reason and the rationale of learning process engineering [5]. It describes the organizational environment of the educational process by indicating the purpose and policy process management. The purpose facet includes two attributes: process and learning. In fact, the construction approaches have been designed for different purposes and try to describe the learning process in different attitudes: descriptive, prescriptive and explanatory. The learning purpose defines the level acquired by constructing the learning process. According the Bloom's Taxonomy [11], the different levels are knowledge, comprehension, application, analysis, synthesis, and evaluation.

TABLE IV. SCENARIO VIEW

Facet	Definition	
	Attribute	***Values***
Purpose	Process	ENUM {Prescriptive, Descriptive, Explanatory}
	Learning	ENUM {Knowledge, Comprehension, Application, Analysis, Synthesis, Eval-
Policy	Reuse	Boolean
	Evolving	Boolean
	Assessment	Boolean

The second attribute identifies three policies of process management namely evolving, reuse and assessment. This attribute supports the validation of construction process quality. More- over, since the learning situations change and evolve over time, it is essential that the learning process construction supports these evolutions [5]. As with any process development, the reuse and the assessment are important, which can occur at any stage of learning process construction and at any level of abstraction and may involve any element of design and/or implementation.

IV. EDUCATIONAL PROCESSES CONSTRUCTION IN ITS

A. Study Scope

Several intelligent tutoring systems have been reported in the literature. Before applying our framework of educational process construction to these systems, Table VI presents a brief description of selected systems. In fact, this selection is based on

- Various didactic domains;
- Different techniques of artificial intelligence;
- Several countries from different continents and,
- Number and quality of related publications

TABLE V. SELECTED ITS

Intelligent Tutoring Systems			
ID	***Name***	***Didactic Domain***	***Country***
ITS 1	ITS-C	Linguistics	Spain
ITS 2	PEGASE	Virtual reality	France
ITS 3	CIRCISM-Tutor	Medicine	USA
ITS 4	Bits	Programming	Canada
ITS 5	SQL-Tutor	Data Base	New Zealand

B. Systems Positions according to the Framework

1) ITS-C: The Intelligent Tutoring System based on Competences (ITS-C) extends an ITS by linking the latter and the pedagogical model based on Competency based Education [12]. It adopts the Computerized Adaptive Tests (CAT) as common tools for the diagnosis process [12].

a) Nature View: The system adopts the learner-directed orientation by using the indirect and the independent methods (see Table VI). The system is used to achieve the intellectual skill and the cognitive strategy by applying the using and the finding performance. In fact, this system supports a tactic and linear process by applying the structuring and the tuning modes.

TABLE VI. NATURE VIEW OF ITS1

Facet	**Definition**	
	Attribute	***Values***
Pedagogy	Orientation	{Learner- Directed}
	Method	{Indirect, Independent Study}
	Correlation	{Not Considered}
Learning	Performance	{Use, Find}
	Mode	{Structuring, Tuning}
	Outcome	{Intellectual Skill, Cognitive Strategy}
Proces	Type	{Tactic}
	Form	{Linear}

b) Description View: By adopting the learner model, the pedagogical model and the domain model, the system adopts a standard and formal notation of learning process (see Table VII). This notation is used to de- scribe the required concepts. By using the aggregation, the system uses a description of activities sequence to cover the desired product.

TABLE VII. DESCRIPTION VIEW OF ITS1

Facet	**Definition**	
	Attribute	***Values***
Notation	Type	{Standard}
	Form	{Formal}
	Content	{Concept}
	Models	{Domain Model, Learner Model, Pedagogical Model}
Level	Performance	{Sequence}
	Mode	{Aggregation}
	Outcome	{Product}

c) Design View: By respecting an uncertain and an evolved context, the system supports the contingency approach. This construction approach applies the assembly methods by adopting problem solving support technique (see Table VIII)

The system adopts a network-based model of optimization by defining an exact method. This method is specified by using different rules. Moreover, the system supports a flexible guidance by using discovery methods to provide the suitable activity. For that, these methods consider cognitive state. To offer a structure-based adaptation, the system uses the method aptitude- treatment interaction by adopting the technique content discovery and assembly. This technique considers the prior knowledge to satisfy the position system initiated adaptivity with pre-information to the user about the change.

TABLE VIII. DESIGN VIEW OF ITS1

Facet	**Definition**	
	Attribute	***Values***
Context	Type	{Uncertain}
	Form	{Evolved}
Construction	Approach	{Contingency}
	Method	{Assembly}
	Technique	{Problem Solving Support}
Optimization	Model	{Network-Based}
	Method	{Exact}
	Technique	{Rule-Based}
	Parameters	{Not Defined}
Guidance	Type	{Flexible}
	Method	{Discovery}
	Parameters	{Cognitive State}
	Outcome	{Activity}
Adaptation	Dimension	{Structure}
	Position	{System Initiated Adaptivity with pre-information to the user about the change}
	Method	{Aptitude-Treatment Interaction}
	Technique	{Content Discovery and Assembly}
	Parameters	{Prior Knowledge})

d) Scenario View: To achieve the evaluation level, the system supports a prescriptive process of learning. The reuse and the assessment of process are considered (see Table IX).

TABLE IX. SCENARIO VIEW OF ITS1

Facet	Definition	
	Attribute	***Values***
Purpose	Process	{Prescriptive}
	Learning	{Evaluation}
Policy	Reuse	{True}
	Evolving	{False}
	Assessment	{True}

2) PEGASE: The PEdagogical Generic and Adaptive SystEm (PEGASE) is used to instruct the learner, and to assist the instructor. This system emits a set of knowledge (actions carried out by the learner, knowledge about the field, etc.) which PEGASE uses to make informed decisions [13].

a) Nature View: The system adopts the learner-directed and the teacher-learner negotiated pedagogical orientations by using the experiential pedagogical methods (see Table X). By considering the pedagogical correlation, the system is used to reach the desired cognitive strategy. This system supports a strategic and linear process of learning by applying the structuring and the tuning modes. These modes are adopted to achieve the using and the finding performance.

TABLE X. NATURE VIEW OF ITS2

Facet	Definition	
	Attribute	***Values***
Pedagogy	Orientation	{Learner-Directed, Teacher-Learner Negotiated}
	Method	{Experiential}
	Correlation	{Considered}
Learning	Performance	{Use, Find}
	Mode	{Structuring, Tuning}
	Outcome	{Cognitive Strategy}
Proces	Type	{Strategic}
	Form	{Linear}

b) Description View: The system applies a standard and formal notation of learning process by using the learner model, the domain model and the pedagogical model (see Table XI). This notation is used to describe the appropriate procedure. By using the aggregation, the system uses a description of activities sequence to consider the context.

TABLE XI. DESCRIPTION VIEW OF ITS2

Facet	Definition	
	Attribute	***Values***
Notation	Type	{Standard}
	Form	{Formal}
	Content	{Procedure}
	Models	{Domain Model, Learner Model, Pedagogical Model}
Level	Performance	{Sequence}
	Mode	{Aggregation}
	Outcome	{Context}

c) Design View: By considering a certain and an evolved context of construction, the system supports the on-the- fly approach. This construction approach applies ad hoc methods by adopting intelligent solution analysis technique (see Table XII).

By adopting the achieved learning intention, the system applies exact method of optimization, by defining different rules. In addition, the system supports a flexible guidance by using discovery and expository methods to provide the suitable activity. For that, these methods con- sider cognitive state. To support a presentation-based adaptation, the system uses the method constructivistic-collaborative by applying the technique adaptive collaboration support. This technique considers the learning history to offer the position system initiated adaptivity with pre- information to the user about the change.

TABLE XII. DESIGN VIEW OF ITS2

Facet	Definition	
	Attribute	***Values***
Context	Type	{Certain}
	Form	{Evolved}
Construction	Approach	{On-The-Fly}
	Method	{Ad hoc}
	Technique	{Intelligent Solution Analysis}
Optimization	Model	{Not Defined}
	Method	{Exact}
	Technique	{Rule Based}
	Parameters	{Achieved Learning Intention}
Guidance	Type	{Flexible}
	Method	{Discovery, Expository}
	Parameters	{Cognitive State}
	Outcome	{Activity}
Adaptation	Dimension	{Presentation}
	Position	{System Initiated Adaptivity with pre-information to the user about the change}
	Method	{Constructivistic-Collaborative}
	Technique	{Adaptive Collaboration Support}
	Parameters	{Learning History}

d) Scenario View: To achieve the evaluation level, the system supports a prescriptive process of learning. The reuse and the assessment of process are considered (see Table XIII).

TABLE XIII. SCENARIO VIEW OF ITS2

Facet	Definition	
	Attribute	***Values***
Purpose	Process	{Descriptive, Explanatory}
	Learning	{Comprehension, Application}
Policy	Reuse	{False}
	Evolving	{True}
	Assessment	{True}

3) CIRCSIM-Tutor: CIRCSIM-Tutor is an intelligent tutoring system for teaching the baroreceptor reflex mechanism of blood pressure control to first-year medical students [14].

a) Nature View: The system adopts the teacher-learner negotiated orientation by using the indirect and the interactive pedagogical methods (see Table XIV). The system is used to reach the desired cognitive strategy and the suitable intellectual skill. For that, the system supports a tactic and linear process of learning by applying the structuring mode. This learning mode is applied to achieve the different types of performance.

TABLE XIV. NATURE VIEW OF ITS3

Facet	Definition	
	Attribute	***Values***
Pedagogy	Orientation	{Teacher-Learner Negotiated}
	Method	{Indirect, Interactive}
	Correlation	{Not considered}
Learning	Performance	{Remember, Use, Find}
	Mode	{Structuring}
	Outcome	{Intellectual Skill, Cognitive Strategy}
Proces	Type	{Tactic}
	Form	{Linear}

b) Description View: The system applies a proprietary and informal notation of learning process by using the learner model and the domain model (see Table XV). This notation is adopted to describe the required procedure. By using the aggregation, the system supports an activity description of learning process.

TABLE XV. DESCRIPTION VIEW OF ITS3

Facet	Definition	
	Attribute	***Values***
Notation	Type	{Proprietary}
	Form	{Informal}
	Content	{Procedure}
	Models	{Domain Model, Learner Model}
Level	Performance	{Activity}
	Mode	{Aggregation}
	Outcome	{Activity}

c) Design View: The system adopts the on-the-fly approach by applying ad hoc methods. This approach supports a certain and stable context of construction by applying problem solving support technique (see Table XVI).

The system supports an exact method by specifying different rules of optimization. Moreover, it adopts a flexible guidance by using collaborative methods to provide the suitable activity. For this purpose, these methods consider the cognitive state. In addition, the system provides a content-based adaptation by using the method constructivistic-collaborative. By using the technique adaptive collaboration support, the method adopts the prior knowledge to satisfy the position system initiated adaptivity with pre-information to the user about the change.

TABLE XVI. DESIGN VIEW OF ITS3

Facet	Definition	
	Attribute	***Values***
Context	Type	{Certain}
	Form	{Stable}
Construction	Approach	{On-The-Fly}
	Method	{Ad hoc}
	Technique	{Problem Solving Support}
Optimization	Model	{Not Defined}
	Method	{Exact}
	Technique	{Rule Based}
	Parameters	{Not Defined}
Guidance	Type	{Flexible}
	Method	{Collaborative}
	Parameters	{Cognitive State}
	Outcome	{Activity}
Adaptation	Dimension	{Content}
	Position	{System Initiated Adaptivity with pre-information to the user about the change}
	Method	{Constructivistic-Collaborative}
	Technique	{Adaptive Collaboration Support}
	Parameters	{Prior Knowledge}

d) Scenario View: The system supports a prescriptive process of learning to achieve the comprehension level. For that, the assessment and the reuse of process are considered (see Table XVII).

TABLE XVII. SCENARIO VIEW OF ITS3

Facet	Definition	
	Attribute	***Values***
Purpose	Process	{Prescriptive}
	Learning	{Comprehension}
Policy	Reuse	{True}
	Evolving	{False}
	Assessment	{True}

4) BITS: The Bayesian Intelligent Tutoring System (BITS) is a web-based intelligent tutoring system for Computer Programming using Bayesian technology [15].

a) Nature View: By considering the learner-directed orientation, the system adopts the direct pedagogical methods (see Table XVIII). The system is used to reach the desired cognitive strategy, the required intellectual skill and the appropriate verbal information. For this purpose, the system supports a tactic and linear process of learning. This process applies the different learning modes to achieve the using and the remembering performance.

TABLE XVIII. NATURE VIEW OF ITS4

Facet	Definition	
	Attribute	***Values***
Pedagogy	Orientation	{Learner-Directed}
	Method	{Direct}
	Correlation	{Not considered}
Learning	Performance	{Remember, Use}
	Mode	{Accretion, Structuring}
	Outcome	{Verbal Information, Intellectual Skill, Cognitive Strategy}
Proces	Type	{Tactic}
	Form	{Linear}

b) Description View: The system applies a standard and formal notation of learning process by using the learner model and the domain model (see Table XIX). This notation is used to describe the required concepts. The system uses an aggregation or primitive description of learning activity.

TABLE XIX. DESCRIPTION VIEW OF ITS4

Facet	Definition	
	Attribute	***Values***
Notation	Type	{Standard}
	Form	{Formal}
	Content	{Concept}
	Models	{Domain Model, Learner Model}
Level	Performance	{Activity}
	Mode	{Aggregation, Primitive}
	Outcome	{Activity}

c) Design View: The system supports contingency approach to consider uncertain and stable context. This approach applies the instantiation methods of construction by adopting curriculum sequencing tech- nique (see Table XX).

By considering the network-based model, the system adopts exact method of optimization by defining different rules. Moreover, the system supports a flexible guidance by using expository methods to offer the suitable resource. To this end, these methods consider cognitive state. To support a structure and content-based adaptation, the system uses the method aptitude-treatment interaction by applying the technique adaptive course delivery. This technique adopts the learning goal and the prior knowledge to offer the position user selection adaptation for system suggested features.

TABLE XX. DESIGN VIEW OF ITS4

Facet	Definition	
	Attribute	***Values***
Context	Type	{Uncertain}
	Form	{Stable}
Construction	Approach	{Contingency}
	Method	{Instantiation}
	Technique	{Curriculum Sequencing}
Optimization	Model	{Network-Based}
	Method	{Exact}
	Technique	{Rule Based}
	Parameters	{Not Defined}
Guidance	Type	{Flexible}
	Method	{Expository}
	Parameters	{Cognitive State}
	Outcome	{Resource}
Adaptation	Dimension	{Content, Structure}
	Position	{User Selection Adaptation for system suggested features}
	Method	{Aptitude-Treatment Interaction}
	Technique	{Adaptive Course Delivery}
	Parameters	{Learning Goal, Prior Knowledge})

d) Scenario View: The system supports a descriptive process of learning to achieve the knowledge and the compre- hension level. For this purpose, the reuse and the evolving of process are adopted by the system (see Table XXI).

TABLE XXI. SCENARIO VIEW OF ITS4

Facet	Definition	
	Attribute	***Values***
Purpose	Process	{Descriptive}
	Learning	{Knowledge, comprehension}
Policy	Reuse	{True}
	Evolving	{True}
	Assessment	{False}

5) SQL-Tutor: The SQL-Tutor is a problem-solving environment intended to complement classroom instruc- tion, and we assume that students are already familiar with database theory and the fundamen- tals of SQL [16]. This ITS adopts the constraint-based modeling as learner modeling approach.

a) Nature View: The system adopts the learner-directed orientation to support the independent study and the in- direct pedagogical methods (see Table XXII). The system is used to reach the desired cognitive strategy by supporting a tactic and linear process of learning. The latter applies the structuring and the tuning modes to achieve the using and the finding performance.

TABLE XXII. NATURE VIEW OF ITS5

Facet	Definition	
	Attribute	***Values***
Pedagogy	Orientation	{Learner-Directed}
	Method	{Indirect, Independent Study}
	Correlation	{Not considered}
Learning	Performance	{Use, Find}
	Mode	{Structuring, Tuning}
	Outcome	{Cognitive Strategy}
Proces	Type	{Tactic}
	Form	{Linear}

b) Description View: The system adopts a proprietary and informal notation of learning process by using the learner model, the pedagogical model and the domain model (see Table XXIII). This notation is adopted to describe the appropriate procedure.

TABLE XXIII. DESCRIPTION VIEW OF ITS5

Facet	Definition	
	Attribute	***Values***
Notation	Type	{Proprietary}
	Form	{Informal}
	Content	{Procedure}
	Models	{Domain Model, Learner Model, Pedagogical Model}
Level	Performance	{Activity}
	Mode	{Primitive}
	Outcome	{Product}

c) . Design View: The system adopts on-the-fly approach to consider certain and stable context of construction. In fact, it supports ad hoc methods by applying intelligent solution analysis technique (see Table XXIV).

The system adopts an exact method by specifying different rules. In addition, the system adopts a flexible guidance by using discovery methods to provide the appropriate resource. To this end, these methods adopt a cognitive state. Furthermore, the system offers a content-based adaptation by using the method aptitude-treatment interaction. This method adopts the technique adaptive interaction by considering the learning history. In fact, this method is used to support the position of system initiated adaptivity with pre-information to the user about the change.

TABLE XXIV. DESIGN VIEW OF ITS5

Facet	Definition	
	Attribute	***Values***
Context	Type	{Certain}
	Form	{Stable}
Construction	Approach	{On-The-Fly}
	Method	{Ad hoc}
	Technique	{Intelligent Solution Analysis}
Optimization	Model	{Network-Based}
	Method	{Exact}
	Technique	{Rule Based}
	Parameters	{Not Defined}
Guidance	Type	{Flexible}
	Method	{Discovery}
	Parameters	{Cognitive State}
	Outcome	{Resource}
Adaptation	Dimension	{Content}
	Position	{System Initiated Adaptativity with pre-information to the user about the change}
	Method	{Aptitude-Treatment Interaction}
	Technique	{Adaptive Interaction}
	Parameters	{Learning History})

d) Scenario View: The system adopts an explanatory process of learning to achieve the comprehension and the ap- plication level. For this purpose, the assessment and the evolving of process are supported by the system (see Table XXV).

TABLE XXV. SCENARIO VIEW OF ITS5

Facet	Definition	
	Attribute	***Values***
Purpose	Process	{Explanatory}
	Learning	{Comprehension, Application}

Facet	Definition	
	Attribute	***Values***
Policy	Reuse	{False}
	Evolving	{True}
	Assessment	{True}

C. Results and Discussion

The framework analysis identifies the following main drawbacks of existing adaptive construction approaches. It should be noted that construction approaches are not sufficiently automated for deriving automatically the objectives achievement from the strategic process.

For nature view, the different systems do not adopt all pedagogical methods to achieve the different types of process outcomes. In fact, the outcomes attitude and motor skill are not considered. In addition, the different systems support tactic and linear process of learning.

For description view, most of the systems try to support the learner model, the domain model and the pedagogical model by applying proprietary notation, which can present some problem of standardization. This notation is not used to specify a process description by respecting the activity level. In fact, the context coverage is not supported by the learning process description.

For design view, the uncertain and the evolved context of construction are not supported by the systems. The different on-the-fly approaches do not support a combined construction techniques of learning process to implement the instantiation methods. Most of these approaches are not implemented by applying a MAP based model of optimization. For that, the educational intentions are not defined as optimization parameters. The construction approaches themselves are not sufficiently guided by considering the different methods and parameters. In fact, the flexible guidance is most used by the expository methods. However, this flexible guidance doesn't adopt the intention as outcome by considering the learning style and the teaching style.

Finally, the different systems do not apply micro methods by respecting system information to support process adaptation.

Therefore, the comparative study identifies the following limits:

- The adaptive ability of most of these ITS is limited to the content. In fact, the process-oriented adaptation is not supported by the different systems.
- The consideration of the educational process as a strategic and not linear one is limited.
- The flexible guidance of educational processes doesn't adopt the intention as outcome by considering the learning style and the teaching style.
- The different systems did not respect the evolving process policy.

V. CONCLUSION AND FUTURE WORKS

This paper presents a new multi-levels view of educational process definition. This view is introduced to present a comparison framework, which has allowed identification of the characteristics and drawbacks of some existing construction approaches supported by ITS. This framework considers adaptive learning process engineering from four different but complementary points of view (Nature, Description, Design and Scenario). Each view allows us to capture a different aspect of learning process engineering.

In fact, the framework is applied to respond to the following purposes: to have an overview of the existing construction approaches, to identify their drawbacks, and to analyze the possibility of proposing a better approach.

The analysis of existing approaches definition identifies the lack of an adaptive and guided construction of educational processes that adopt the new technologies of artificial intelligence. It should be noted that the majority of ITS do not respect the correlation between pedagogical and learning process. Moreover, the learner's motivation is not considered.

REFERENCES

[1] I. Bayoudh Saâdi, W. Bayounes, and H. Ben Ghezala, " Educational processes'guidance based on evolving context prediction in intelligent tutoring systems",Universal Access in the Information Society, vol. 8, no.68, pp 1-24. 2019.

[2] W. Bayounes, I. B. Saâdi, Kinshuk, and H. B. Ghézala, "An Intentional Model for Ped- agogical Process Guidance Supported by an Adaptive Learning System", in Proc. 23rd IBIMA Conference, Valencia, Spain, 2014, pp. 1211–1227.

[3] W. Bayounes, I. B. Saâdi, Kinshuk, and H. B. Ghézala, "An Intentional Model for Learning Process Guidance in Adaptive Learning System", in Proc. 22nd IBIMA Conference, Rome, Italy, 2013, pp. 1476–1490.

[4] C. Rolland, "A Comprehensive View of Process Engineering", in Proc. 10th CAISE Conference, Pisa, Italy, C. T. Spring Verlag Percini, Ed., 1998, pp. 1–24.

[5] W. Bayounes, I. B. Saâdi, and H. B. Ghézala, "Towards a Framework Definition for Learning Process Engineering Supported by an Adaptive Learning System", in Proc. ICTEE Conference, Amritapuri, India, 2012, pp. 366–373.

[6] M. D. Merrill, "Component display theory", IInstructional Design Theories And Models: An Overview Of Their Current Status, vol. 1, pp. 282–333, 1983.

[7] D. E. Rumelhart and D. A. 2, "Accretion, Tuning, and Restructuring: Three Modes of Learning", in Semantic Factors in Cognition, W. Cotton and R. Klatzky, Eds., Hillsdale, NJ: Erlbaum, 1978, pp. 37–53.

[8] A. Patel and K Kinshuk, "Intelligent Tutoring Tools in a ComputerIntegrated Learning Environment for Introductory Numeric Disciplines", Innovations in Education & Training International, vol. 34, no. 3, pp. 200–207, 1997.

[9] V. M. Garcia-barrios, F. Mödritscher, and G. Christian, "The Past, the Present and the Future of adaptive E-Learning: An Approach within the Scope of the Research Project AdeLE", in Proc. ICL Conference, Villach, Austria, 2004, pp. 1–9.

[10] A. Paramythis and S Loidl-Reisinger, "Adaptive learning environments and e-learning standards", Electronic Journal of Elearning, vol. 2, no. 1, pp. 181–194, 2004.

[11] M. Forehand, "Bloom's Taxonomy", Emerging Perspectives on Learning, Teaching, and Tech- nology, p. 12, 2012.

[12] M. Badaracco and L. Martínez, "A fuzzy linguistic algorithm for adaptive test in Intelligent Tutoring System based on competences", Expert Systems with Applications, vol. 40, no. 8, pp. 3073–3086, 2013.

[13] C Buche, C Bossard, R Querrec, and P Chevaillier, "{PEGASE}: A Generic and Adaptable Intelligent System for Virtual Reality Learning Environments", IJVR, vol. 9, no. 2, pp. 73– 85, 2010.

[14] M. Glass, “Processing language input in the CIRCSIM-Tutor intelligent tutoring system”, Artificial Intelligence in Education, pp. 210–221, 2001.

[15] C. Butz, S. Hua, and R. B. Maguire, “Bits: a Bayesian Intelligent Tutoring System for Computer Programming”, Wccce ’04, pp. 179–186, 2004.

[16] A. Mitrovic, B. Martin, and M. Mayo, “Multiyear Evaluation of SQL-Tutor : Results and Experiences”, pp. 1–54, 2000.

A medical decision support system for cardiovacsular disease based on ontology learning

Samia Sbissi
SMART Laboratory,
Tunis University, Tunisia
School of Engineering,
ESPRIT University, Z.I. Chotrana II. B.P,
160-2083, Tunis, Tunisia
Email: samia.sbissi@esprit.tn

Mariem Mahfoudh
MIRACL Laboratory,
University of Sfax, Tunisia
ISIGK Laboratory,
University of Kairouan, Tunisia
Email: mariem.mahfoudh@gmail.com

Said Gattoufi
SMART Laboratory,
Tunis University, Tunisia
Email: algattoufi@yahoo.com

***Abstract*—This paper presents our decision support system "CardioSAD" that aims to assist cardiologists to make relevant decisions for patients who are at risk cardiovascular disease. The idea consists to analyse clinical practice guidelines (documents containing recommendations and medical knowledge) to enrich and exploit an existing ontology. The enrichment process is driven by the ontology learning task. We start by pre-processing the text and extracting the relevant concepts. Then we enrich the ontology with OWL DL axioms and SWRL rules. These rules will be inferred for suggesting appropriate recommendations to the doctors.**

1. Introduction

Following medical developments and the evolution of disease treatments, which are experiencing exponential growth, is a nontrivial task. This makes it difficult for a medical expert to assimilate new documents and make the appropriate decisions. These documents are called clinical practice guidelines and contain a set of recommendations and knowledge that make to appropriate decisions and improve the quality of cares. Cardiovascular diseases are the first cause of loss of Disability Adjusted Life Years (DALYs) in developed countries (accounting for more than 23% of the total number), higher than neoplasia (17,9%) and neuropsychiatric conditions (16,5%) [1]. For instance, symptoms of cardiovascular disease as dissection aortic are not easily observable by an individual. DPSS, supported by medical sensors, machine learning algorithms, and computing facilities, is used for the proper assessment of the disease [2]. To Recover relevant information and transform text in knowledge interpreting by the machine, the techniques of natrural language processing, specially semantic parsing, are highly needed. These techniques consist of producing a formal representation and a meaning of a text written in natural language [3] which can be used for example in reasoning and inference or also to determine whether two texts are in relation to involvement [4].
The automatic retrieval of knowledge from text has been at the heart of several areas of research as the volume of data available has increased dramatically from year to year in the form of corpus texts [5], [6].
Ontology plays a key role in representing hidden knowledge in this text and makes it computerized and comprehensible. Ontologies can be created by extracting relevant information from text using a process called ontology enrichment. The corresponding terms and synonyms are automatically transformed into concepts. Then the taxonomic and non-taxonomic relationships between these concepts are generated. Finally, axiom patterns are instantiated and general axioms are extracted from unstructured text. This whole process is known as ontology learning.
In order to develop the decision support system, we need to model the knowledge of this disease through ontology. One of the ontologies found close to our domain is ontology CVDO [1]. It is an OWL ontology, designed to describe entities related to cardiovascular disease. This ontology will be enriched and based on the knowledge that is generally described in specialized documents called clinical practice guidelines [7]. They consist of a set of rules and recommendations. Example: `In patients with abdominal aortic diameter of 25-29 mm, new ultrasound imaging should be considered 4 years later`).
Our goal is to transform these recommendations into logical forms in the form of semantic web rule language (SWRL).
SWRL [8] is a semantic web language directly integrated with OWL ontologies (ontology web language). The transformation of recommendations into SWRL rules will be driven by the process of ontology learning. We integrate the results to enrich the CVDO ontology and

1. http://purl.bioontology.org/ontology/CVDO

use it in inference tasks in order to develop a decision support system for the cardiologists.
The remainder of the article is organized as follows. Section 2 provides a detailed study of relevant works. Section 3 describes our proposed approach. Section 4 presents the implementation and results. Finally, Section 5 concludes the paper.

2. Related works

Decision support systems exist in many application areas where human experts need help in retrieving and analyzing information such as: financial management, ship routing, medical diagnosis, climate change, etc. [9]. In our work, we are interested in decision support system dedicated to the medical field. Adoptiing decision support systems in the diagnosis and management of chronic diseases, such as diabetes [10], the cancer [11], the cardiac disease [12], and high blood pressure [13] has played an important role in key health care organizations. These medical systems contribute to the improvement of the quality of care [13]. Diagnostic support systems can be classified into systems using structured knowledge, systems using unstructured knowledge and systems using medical decision formulas (rules or algorithms). Among the first diagnostic systems based on structured knowledge is MYCIN [14]. It is one of the first computerized expert systems to facilitate the diagnosis and treatment of certain blood infections [15]. It was based on a knowledge base in the form of decision-making rules and an inference engine [16]. In 2012, the authors of [17] used semantic web techniques to model good clinical practice guidelines (GBPC). These guidelines have been coded as a set of rules (through domain ontology) used by SADM to generate specific recommendations and assess risks in breast cancer patients. In 2015, authors [18] proposed a SADM based on the use of ontology and SWRL rules. In making a decision, the system considers patient information, signs and symptoms, laboratory tests, and diabetes risk factors. However the incompleteness of the system poses a problem since the authors have modelled only an ontology and rules, but there is no interface used by the users.
Ontology-based referral systems could be used for cardiovascular disease prognosis and prevention. This could help physicians predict a patient's worsening health status by reviewing the patient's medical history. A system was subsequently developed in 2016 by [19]. It is a hybrid system based on ontology and the techniques of "Machine Learning" (ML). It performs a complex job classification and helps physicians automatically distinguish patients with chest pain from others. Two criteria are used in the classification which are disease risk factors (such as age, sex..), and laboratory test results using ML techniques for classification (SVM, Logic Regression LR, Decision Tree DT, Backward Selection BS, Forward Selection FS,..).
The table 1 is a summary table of the approaches presented above.

In recent years, there has been a surge in research on knowledge acquisition from text and other sources, which is to a great extent due to a renewed interest in knowledge-based techniques in the context of the Semantic Web endeavor. There is a large body of research on automating the ontology learning using natural language processing (NLP). Moreover, ontological knowledge is often stated explicitly or implicitly within the text, and these reference documents serve as important resources for ontology learning. It can be processed in three ways: manual, semi-automatic and automatic. There are several systems proposed to deal with ontology learning: Text-to-Onto [22], due to its significance to a wide range of researchers as well as practitioners for purposes ranging from e-Learning [23] and e-Government. OntoGain [24], is included due to its recency to the community, which will act as a yardstick to examine the progress of ontology learning systems over the past ten years. Authors in [25] have built an ontological learning system by collecting evidence from heterogeneous sources in a statistical approach. In the medical field, there are a lot of non-taxonomic relationships, such as symptoms and ethologies of diseases, indications of medicines, aliases of diseases or medicines, etc. Among them, ontology learning from this type of text play a big role using statistical and linguistic methods. [26] propose a skip gram model implemented in the word2vec system. The main idea is that words with similar contexts should have similar meanings. For example, if we see the two sentences "the patient complained of aortic dissection symptoms" and "the patient reported aortic dissection symptoms", we might infer that "complained" means the same thing as "reported". As a result, these two words should be close in the representation space. [27] learn embedding from unstructured medical corpora crawled from PubMed, Merck Manuals, Medscape and Wikipedia.

3. Proposed approach

In order to assist the doctors in the decision-making process and offer recommendations for a given patient, we have proposed a system that we have named "CardioSAD" illustrated in figure 1.

In order to build our system "CardioSAD", we have passed by the following steps:

- Construction of the recommendation corpus
- Evolve existing ontology "CVDO" by recommendation text
- Build our system "CardioSAD" then request and infer ontology to have the best recommendations.

TABLE 1: Medical decision support system using ontologies

The system **COMET** by Abidi and al. [17]	Provide personalized recommendations for patients with cardiovascular disease	• System used for small data. • Requires multiple user interventions. • Causes contradictory results.
The system developed by **Alharbi and al.** [18]	Uses an ontology, SWRL rules, information about the patient, signs and symptoms, laboratory tests and diabetes risk factors.	• Incomplete system. • No interface. • Ontology is not complete which alters the diagnostic quality
The system developed by **Farooq and al.** [19]	An hybrid system for decision support based on semantic web techniques and machine learning	• The decision for diagnosis depends on specific risk factors that need to combine several ML techniques.
The system developed by **Alexendre** in 2015 [20]	A SADM which assist the physician in cardiovascular disease such as chronic polypathologies.	• Patient profiles are organized hierarchically. • Considers only two disease risk factors for making the decision.

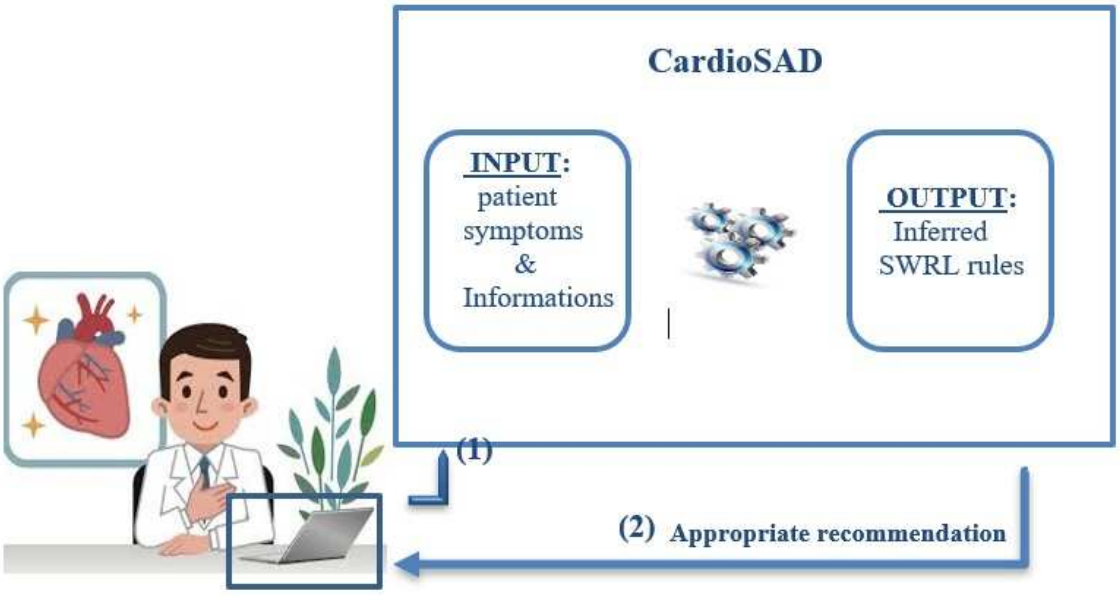


Figure 1: The "CardioSAD" overview

3.1. Construction of recommendation corpus

Our recommendations corpus is developed by the European Society of Cardiology (ESC) are published on its website [2]. The recommendations are designed to help health professionals to make decisions about patients. The table 2 contains an extract from the recommendations.

3.2. Ontology learning

In this section, we describe the ontology learning process and the construction of SWRL rules that will

2. http://www.escardio.org/guidelinessurveys/ esc-guidelines/about/Pages/rules-writing.aspx

"In patients with acute contained rupture of TAA, urgent repair is recommended." "In cases of aneurysm of the abdominal aorta, duplex ultrasound for screening of peripheral artery disease and peripheral aneurysms should be considered."

TABLE 2: Some rules extract from clinical practice guideline.

be integrated in the CVDO ontology. This ontology will be used in the process of text annotation and ontology learning.
To pre-process the text that represents the recommendations, we have used the following steps:

- tokenization and normalization.
- part-of-speech (POS) tagging.
- lemmatization / stemming / morphological analysis.
- chunking / dependency.

We describe briefly some technical of text preprocessing quoted above.

Tokenization: used to detect sentence as well as word boundaries as mentioned in the following example:

In all patients with AD, medical therapy including pain relief and blood pressure control is recommended.

⇓

['In', 'all', 'patients', 'with', 'AD', ',', 'medical', 'therapy', 'including', 'pain', 'relief', 'and', 'blood', 'pressure', 'control', 'is', 'recommended']

Part-of-speech (POS) tagging is the task of assigning to each token its corresponding part-of-speech, i.e. its syntactic word category such as noun, adjective, verb, etc.

In patients with suspected rupture of the TAA, emergency CT angiography for diagnosis confirmation is recommended.

⇓

[In / IN, patients / NNS, with / IN, suspected / VBN, rupture / NN, of / IN, the / DT, TAA / NNP, ,/„ emergency / NN, CT / NN, angiography / NN, for / IN, diagnosis / NN, confirmation / NN, is / VBZ, recommended / VBN, ./.].

Term/Concept extraction: Several approaches use linguistic method to extract terms and concepts [28], [29], [30]. The text is tagged with parts of speech to extract syntactic structures in a sentence such as noun phrases and verb phrases.

Algorithm 1: Concepts and Properties Identification

Result: concept and properties extracted
domainWordsList, POS Tagge;
Begin
Read the domainWordsList as array:
Foreach word **in** domainWordsList **do**:
Get the word Tag using POS Tagger :
if *wordTag is noun or its subsequent* **then**
  Add word to conceptLis;
  elseif wordTag is verb or its subsequent **then**
  Add word to propertiesLis
else
end
Ignore the current word

3.3. Ontology evolution

The figure 2 describes the different steps of the ontology CVDO evolution. The evolution of the CVDO

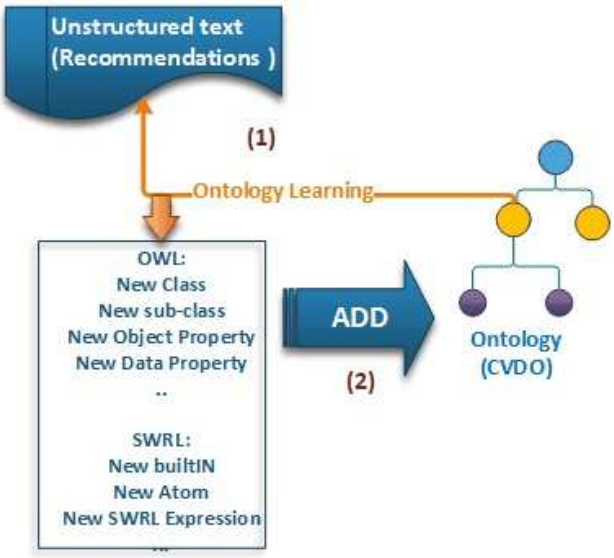


Figure 2: Ontology evolution.

ontology takes place not only by adding elements (concepts, object properties and data properties) resulting from the ontology learning process, but also in adding swrl rules extracted from the clinical practice guidelines. Ontological evolution is achieved through changes.

- **OWL changes:** AddClass, AddSubClass, AddDataProperty, AddObjectProperty, AddDataPropertyAssertion, AddObjectPropertyAssertion, AddIndividual, etc.
- **SWRL changes:** AddAtom, AddClassAtom, SWRLBuiltIn, SWRLExpression, etc.

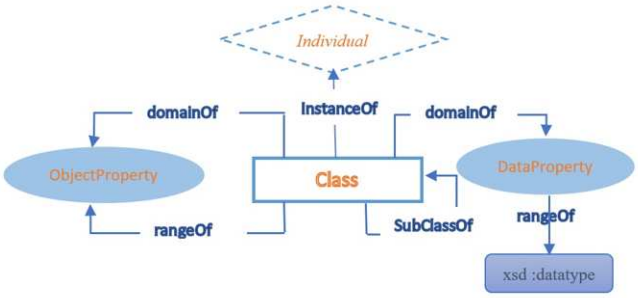


Figure 3: The relationships between the entities of ontology.

3.4. Ontology enrichment and population

Ontology enrichment involves adding or modifying existing ontology performing one or more learning tasks. It is the process of extending an ontology with new concepts, properties and/or terminological axioms. Ontology population is the process of adding new instances in an ontology. This involves adding individual instances of concepts, but also assertions properties related to instances.
Individuals are the instances of the domain. The assertions express the typing of individuals (the individual "surgey" is an instance of the concept "diagnosis") and properties (the triplet <"ALI", "Recommendeddiagnosis", "Surgey"> is an instance of the property "Recommendeddiagnosis").
We are not only interested in simple concepts and taxonomic relationships, but also to SWRL rules automatically extracted from directives medical.
The enrichment process attempts to facilitate text understanding and automatic process textual resources, by the transformation of words to concepts and to relations [31]. It starts by extracting concepts/relationships from the text using a processing language, such as partial speech markup (POS) and segmentation of phrase. The concepts and relationships extracted are then organized in ontology using syntactic and semantic analysis techniques.
The text contains a set of recommendations. The following example presents a recommendation and its treatment.

Identification of concepts and individuals. As it is mentioned in the figure 4 that presents the ontology learning process, some nouns are mentioned as classes and others as individuals. Let see the difference.

In patients with suspected rupture of the TAA, emergency CT angiography for diagnosis confirmation is recommended.
⇓
`Patient` → identified as class

Emergency_CT_angiography → identified as an individual

We applied the algorithm 1 to identify the concept and properties. For the first elements, the noun is considered as a concept.
Each noun will be added the vector of concepts. But after, we will regroup this vector into two elements (concept or individual).
So the question here why patient is identified as a concept(class) and *emergency_CT_angiograhy* as an individual?
To answer this question, we followed the following steps:

- In [31], we used the technique of matching from text to an ontology, to extract relevant terms from text.
- We have used Levenshtein measure to search similarity between text and the ontology:
  - if we identify a correspondence, in this case, we will do nothing.
  - else if there is no correspondence, we will use WordNet ontology to search, semantic similarity. If it exists, we will add this concept as similar concept, we use AddSimilarClass.
  - else we add this noun as a new class (concept).

  In order to ameliorate our results, we integrated other linguistic and syntactic methods. We also use a statistical method, Word2vec.
  The identification of individuals used in the context of this noun. Example aortic dissection is an individual of existing class disease.
  Also, TEVAR or emergency-CT-angiograhy are identified as individuals of the existing class of treatment, because they are in the same context using word2vec.

  Word2vec [26] computes continuous vector representations for large text data sets. It provides high performance for measuring syntactic and semantic similarities.

4. Implementation

Our work takes place within the framework of a cooperation between the hospital "La Rabta" of Tunis and the SMART laboratory (Straegies for Modelling and Artificial inTelligence) and this in order to assist cardiologists in decision-making for patients with aortic dissection. To this end, we have developed a tool with an implementation based on the Java language. The Stanford Core NLP API is used for the preprocessing of the text, to determine the POS marking and the chunked tree in the process of learning and ontology enrichment. To read and evolve ontology, we used the Jena [3] which

3. https://jena.apache.org/

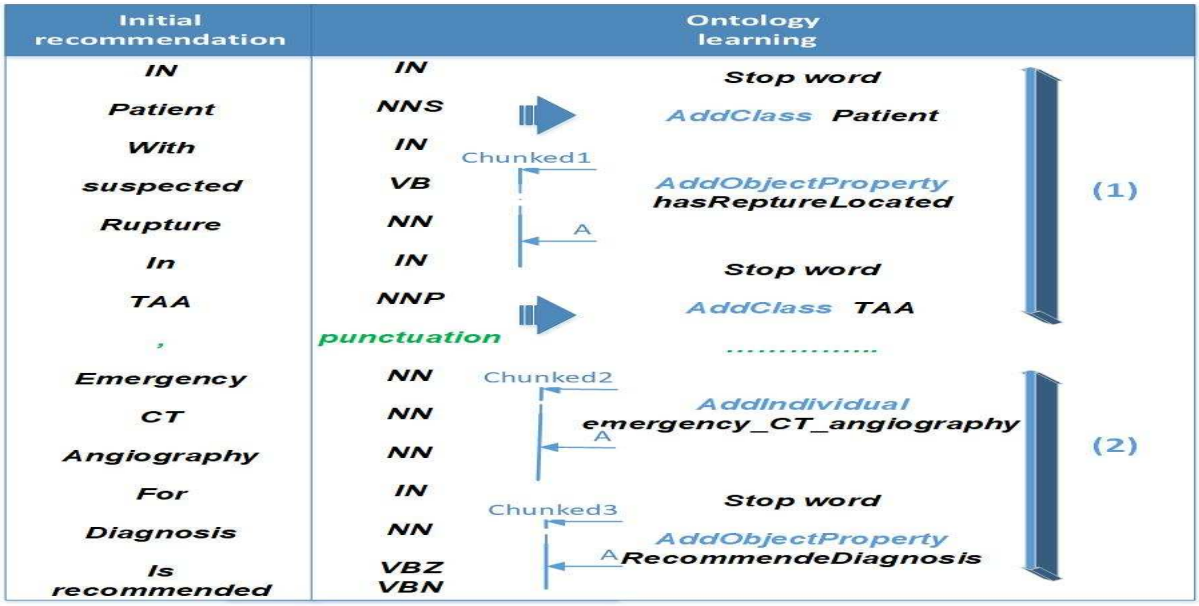


Figure 4: Ontolgy learning.

is a Java API for building semantic Web applications. This framework provides extended Java libraries to help developers develop code that manages RDF, RDFS, OWL and SWRL in accordance with recommendations published by the W3C. Also, we used gensim which is a Python library to test integration words with word2vec.

4.1. Search for correspondences

We proceed through the semantic annotation process in which we search for links between the CVDO ontology and a pre-processed practice guide text clinical. Name concepts are used to produce an extended list of terms equivalent or related persons. Each entry text term can be associated with one or more ontology entities.
To find the similarities, we used [31]:

1) exact matching: identifies identical entities (String) in the text and in the ontology of the domain,
2) morphological correspondence: identifies entities with morphological correspondence,
3) syntactic matching: using the Levenshtein measurement [32],
4) semantic correspondence: identifies synonym relationships with Wordnet ontology.

Table 3 presents the results of the similarity process.

Initially and without recourse to the measures of similarities, we have 4.38% correspondence between text and ontology. It is through the search for the existence of textual terms in the ontology. As a result, we used similarity measures.We have extracted only 30% of the similarities were extracted. The only relationship between the text and ontology is "similar to". This type of relationship is sufficient for an enrichment task. We move on to step 2, that of learning automatic ontology.

TABLE 3: Search for correspondences

	number of links	links(%)
Without similarity	28	4.38%
With similarity	190	30%

4.2. Ontology learning

Linguistics and syntactic analysis is employed head-modifier principal to identify and extract complex terms in which the head of the complex term takes the role of hypernym. X is a hyponym of Y if Y is a type of X. Example a dissection is a hyponym of dissecting aortic. Also other words angiography. We used linguistic features (POS, etc.) and word embedding features (word2vec). The word2vec implementation is extremely simple and provides a high-percentage of relevant concept candidates. On the minus side, candidates suggested by word2vec are (as expected) sometimes even too strongly related to the seed terms, for example, syntactic variations such as plural forms or near-synonyms. Word2vec with bigram produce a better result. We take a random sample of words from the vocabulary of bigrams, just 150 we tested in. Another way to thin this down might be to only pick nouns or only pick the most common words.

In table 4, we have done several iterations of the algorithm. During the first iteration, the algorithm requires the first user intervention to remove all words away from

the domain from the result file. After the third iteration, the algorithm automatically provides a match.

TABLE 4: Word2vec process

	Word2vec (unigram)	Word2vec (bigram)
Iteration1	71.8%	76.2%
IterationN	76.6%	81.8%

One of the advantages of our approach, is the implementation of word2vec which is extremely simple and offers a high percentage of relevant concepts. Among the drawbacks that we met, is that the candidates suggested by word2vec are (as expected) sometimes even too strongly linked to the initial terms, for example syntactic variations such as plural forms or quasi-synonyms. Word2vec with bigram produces a better result.

4.3. SWRL rules and decision making

CVDO ontology is enhanced with SWRL rules to take and deduct good medical decisions (see figure 8). As mentioned, the defined SWRL are essentially of two types.

- The first type allows to deduce the diagnosis necessary for a patient "RecommendedDiagnostic".
- The second makes it possible to deduce the appropriate treatment that must be offered to the doctor "RecommendedTreatment".

We present here some converted recommendation to SWRL rules and we explain them here.

- **R1.**

  In patients with suspected rupture of the TAA, emergency CT angiography for diagnosis confirmation is recommended.

  ⇓

  `cvdo:Patient(?p) ∧ cvdo:hasReptureLocated(?p, "TAA") → cvdo:RecommendedDiagnosis(?p, cvdo:emergency_CT_angiography)`

  ⇒ If the patient had an aneurysm of the thoracic aorta then the doctor recommends an emergency CT angiography in order to confirm the diagnosis.
- **R2.**

  In AD case, surgery may be considered.

  ⇓

  `cvdo:hasDesease(?p, cvdo:AD) ∧ cvdo:Patient(?p) → cvdo:RecommendedTreatment(?p, cvdo:surgey)`

  ⇒ After confirming the diagnosis, if the patient had aortic dissection, an operation will be recommended.
- **R3.**

  In AD case, surgery may be considered.

  ⇓

  `cvdo:hasSymptoms(?p, cvdo:Douleur) ∧ cvdo:hasSymptomsValue(?s, "douleur_thoracique_aigue") ∧ cvdo:Patient(?p) ∧ cvdo:hasSymptoms(?p, cvdo:Hypertension)∧ cvdo:hasSymptomsValue(?s, 14) → cvdo:RecommendedDiagnosis(?p, cvdo:Angioscanner)`

  ⇒ Patients who have acute chest pain and have hypertension, the angiography will be recommended by the doctor for confirmation of the diagnosis.
- **R4.**

  In patients who have maximal aortic diameter 40 mm with Marfan syndrome, surgery is indicated .

  ⇓

  `cvdo:Patient(?p) ∧ cvdo:hasDisease(?p, cvdo:Marfan_syndrome) ∧ cvdo:hasAbdominalDiametre(?p, 40) → cvdo:RecommendedTreatment(?p, cvdo:surgey)`

  ⇒ If the patient had Marfan syndrome, and had the abdominal diameter at about 40 cm, an operation is recommended urgently to save his life.
- **R5.**

  In complicated Type B AD, TEVAR is recommended.

  ⇓

  `cvdo:hasDesease(?p, cvdo:AD) ∧ cvdo:hasDType(?p, cvdo:B) ∧ cvdo:Patient(?p) -> cvdo:RecommendedTreatment(?p, cvdo:TEVAR)`

  ⇒ If the doctor proves the presence of aortic dissection type B in the patient, TEVAR (thoracic endovascular aortic repair) in this case is recommended. TEVAR is an endovascular repair technique consisting of placement of the stent.

60% of recommendations are converted to SWRL rules. The lack of 40% comes back to the fact that we could not convert rules which seems difficult. We take an example of a recommendation that we failed to convert it:

Surveillance is indicated and safe in patients with AAA with a maximum diameter of <55 mm and slow (<10 mm/year) growth.

This can be explained by the fact of using stop words. Slow and growth are identified as stop words that will be ignored.

4.4. CardioSAD system

As part of assisting cardiologists in the decision-making process, we have developed a system that named "CardioSAD" . The goal is to manage as simply possible access to all system features without the doctor's cardiologist will not get lost. We took care in our work to present cardiologists with ergonomic interfaces and simple to use. In what follows, we present some interfaces of our system.

Home page to the cardiologist

Once authentication is successful, the cardiologist will be led to a welcome interface illustrated in the figure 5. Through this interface, the cardiologist can add a patient record. If it does not exist, he can search for a file or update a medical record.

Figure 5: Home page to the cardiologist

Interface to add a patient folder

Adding a new patient can be achieved with the interface presented in the figure 6. The cardiologist captures the patient's data such as: (Patient's name, ID file, age, symptoms,...). By clicking on the "Add" button, a new patient will be added to both the database and ontology.

Interface to ask recommendation

Our system assists the cardiologist to make decision for the diagnosis and treatment of a patient at risk of dissection aortic. Decisions will be presented by recommendations that are inferred and reasoned from SWRL rules. These rules are built from recommendations described by experts in the field. The recommendation interface is shown in Figure 7. So after adding a patient

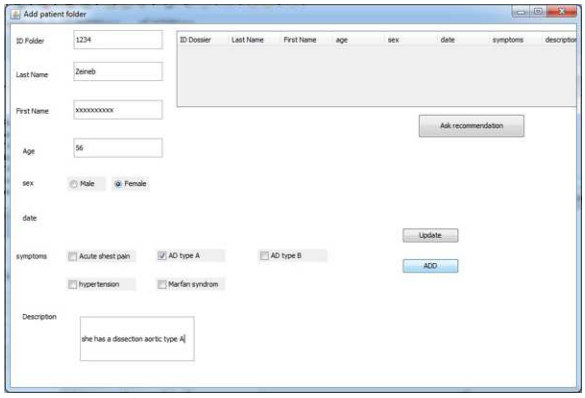


Figure 6: Interface to add a folder of patient

Figure 7: Interface to ask recommendation

or looking for a patient, the cardiologist can click on the "recommendation" button to have the recommendations proposed by the system for a given patient.

Example 1 : Amine is a patient who suffers from acute chest pain "douleur thoracique aiguÃń". The reasoner will recommend the next SWRL rule

$$cvdo : hasSymptoms(?p, cvdo : Douleur)^{c}vdo : hasSymptomsValue(?s, "douleurThoraciqueAigue")^{c}vdo : Patient(?p)\text{->}cvdo : RecommendedDiagnosis(?p, cvdo : Angioscanner)$$

This rule has as a conclusion CT angiography, it is the recommendation for a diagnosis proposed actually by our system illustrated in the figure 8.

Example 2: Mohamed had an aortic dissection type A "AD type A". During the recommendation request process, the following rule will be proposed:

$$cvdo : hasDesease(?p, cvdo : AD)^{c}vdo : hasDType(?p, cvdo : B)^{c}vdo : Patient(?p)\text{->}cvdo : RecommendedTreatment(?p, cvdo : TEVAR))$$

In this case, TEVAR (thoracic endovascular aortic repair) is recommended. Knowing that TEVAR consists

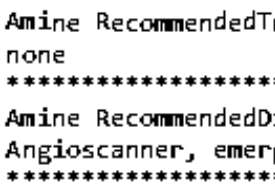


Figure 8: Decision 1 : Angioscanner

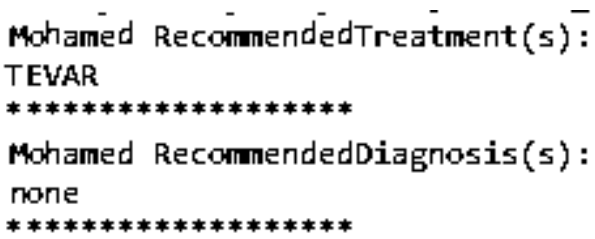


Figure 9: Decision 2 : TEVAR

in placing the stent. The result returned by our system is shown in the figure 9.

Example 3: Zeineb is a patient had aortic dissection type A. The next rule will be triggered which results in surgery.

$$cvdo : hasDesease(?p, cvdo : AD)^c vdo : hasDType(?p, cvdo : A)^c vdo : Patient(?p) -> cvdo : RecommendedTreatment(?p, cvdo : surgey)$$

The goal of emergency surgical treatment is to prevent the fatal complications of intrapericardial aortic rupture. Which is recommended by our system. The result returned by our system is shown in the figure 10.

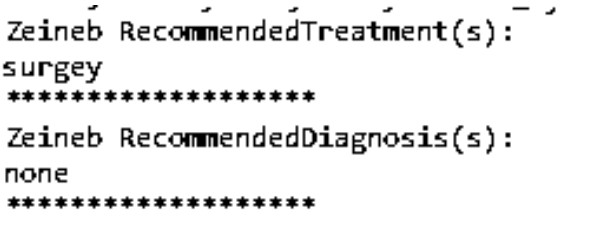


Figure 10: Decision 3 : Surgey

5. Conclusion

In this article, we have proposed a decision support system offering to the cardiologist appropriate recommendations for the diagnosis and treatment of a given patient. Our approach is based on an automatic ontology learning from unstructured text (a Clinical Practice Guideline (CPG)) to enrich an existing ontology. CPG provides a set of recommendations and knowledge to help physicians decide on appropriate health care for patients at risk of cardiovascular disease. The process of ontology learning stars with the analysis of the text using StanfordâĂŹs core NLP. Then, it switches to the extraction of relevant terminology, synonymous with the identification of terms, concepts, construction, concept hierarchy organization, relationships, learning, organisation of hierarchy of relationships and extraction of axioms. To extract the term/concept, we used the Levenshtein measure, the ontology Wordnet, and the statistical method word2vec. For other relationships, we used the segmentation tree and the Hearst model to search for related lines. Once these elements were extracted, we updated the ontology by adding them. The ontological enrichment process treats concepts, OWL axioms and integrates SWRL rules.

References

[1] M. A. Zuluaga, N. Burgos, A. F. Mendelson, A. M. Taylor, and S. Ourselin, "Voxelwise atlas rating for computer assisted diagnosis: Application to congenital heart diseases of the great arteries," *Medical image analysis*, vol. 26, no. 1, pp. 185–194, 2015.

[2] D. Malathi, R. Logesh, V. Subramaniyaswamy, V. Vijayakumar, and A. K. Sangaiah, "Hybrid reasoning-based privacy-aware disease prediction support system," *Computers & Electrical Engineering*, vol. 73, pp. 114–127, 2019.

[3] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, no. 1, pp. 49–79, 2004.

[4] B. Gyawali, A. Shimorina, C. Gardent, S. Cruz-Lara, and M. Mahfoudh, "Mapping natural language to description logic," in *European Semantic Web Conference*. Springer, 2017, pp. 273–288.

[5] R. Upadhyay and A. Fujii, "Semantic knowledge extraction from research documents," in *Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on*. IEEE, 2016, pp. 439–445.

[6] P. Ristoski and H. Paulheim, "Semantic web in data mining and knowledge discovery: A comprehensive survey," *Web semantics: science, services and agents on the World Wide Web*, vol. 36, pp. 1–22, 2016.

[7] R. Erbel, V. Aboyans, C. Boileau, E. Bossone, R. D. Bartolomeo, H. Eggebrecht, A. Evangelista, V. Falk, H. Frank *et al.*, "2014 esc guidelines on the diagnosis and treatment of aortic diseases," *European heart journal*, vol. 35, no. 41, pp. 2873–2926, 2014.

[8] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean *et al.*, "Swrl: A semantic web rule language combining owl and ruleml," *W3C Member submission*, vol. 21, p. 79, 2004.

[9] S. M. Capalbo, C. Seavert, J. M. Antle, J. Way, and L. Houston, "Understanding tradeoffs in the context of farm-scale impacts: An application of decision-support tools for assessing climate smart agriculture," in *Climate Smart Agriculture*. Springer, 2018, pp. 173–197.

[10] P. J. OâĂŹConnor, J. M. Sperl-Hillen, W. A. Rush, P. E. Johnson, G. H. Amundson, S. E. Asche, H. L. Ekstrom, and T. P. Gilmer, "Impact of electronic health record clinical decision support on diabetes care: a randomized trial," *The Annals of Family Medicine*, vol. 9, no. 1, pp. 12–21, 2011.

[11] S. B. Clauser, E. H. Wagner, E. J. A. Bowles, L. Tuzzio, and S. M. Greene, "Improving modern cancer care through information technology," *American journal of preventive medicine*, vol. 40, no. 5, pp. S198–S207, 2011.

[12] R. F. DeBusk, N. Houston-Miller, and L. Raby, "Clinical validation of a decision support system for acute coronary syndromes," *Journal of the American College of Cardiology*, vol. 55, no. 10, pp. A132–E1240, 2010.

[13] K. Farooq, B. S. Khan, M. A. Niazi, S. J. Leslie, and A. Hussain, "Clinical decision support systems: A visual survey," *arXiv preprint arXiv:1708.09734*, 2017.

[14] P. Elkin, M. Peleg, R. Lacson, E. Bernstam, S. Tu, A. Boxwala, R. Greenes, and E. H. Shortliffe, "Toward standardization of electronic guideline representation," *Md Computing*, vol. 17, no. 6, pp. 39–44, 2000.

[15] A. Galopin, J. Bouaud, S. Pereira, and B. Séroussi, "An ontology-based clinical decision support system for the management of patients with multiple chronic disorders." in *MedInfo*, 2015, pp. 275–279.

[16] W. R. Swartout, "Rule-based expert systems: The mycin experiments of the stanford heuristic programming project: Bg buchanan and eh shortliffe,(addison-wesley, reading, ma, 1984); 702 pages, 40.50," 1985.

[17] S. Abidi, J. Cox, M. Shepherd, and S. S. R. Abidi, "Using owl ontologies for clinical guidelines based comorbid decision support," in *System Science (HICSS), 2012 45th Hawaii International Conference on*. IEEE, 2012, pp. 3030–3038.

[18] R. F. Alharbi, J. Berri, and S. El-Masri, "Ontology based clinical decision support system for diabetes diagnostic," in *Science and Information Conference (SAI), 2015*. IEEE, 2015, pp. 597–602.

[19] K. Farooq and A. Hussain, "A novel ontology and machine learning driven hybrid cardiovascular clinical prognosis as a complex adaptive clinical system," *Complex Adaptive Systems Modeling*, vol. 4, no. 1, p. 12, 2016.

[20] B. Séroussi, A. Galopin, M. Gaouar, S. Pereira, and J. Bouaud, "Using therapeutic circles to visualize guideline-based therapeutic recommendations for patients with multiple chronic conditions: A case study with go-dss on hypertension, type 2 diabetes, and dyslipidemia." *Studies in health technology and informatics*, vol. 245, pp. 1148–1152, 2017.

[21] H. Hjelm and M. Volk, "Cross-language ontology learning," pp. 272–297, 2011.

[22] P. Cimiano, A. Mädche, S. Staab, and J. Völker, "Ontology learning," pp. 245–267, 2009.

[23] J.-P. Hatala and J. George Lutta, "Managing information sharing within an organizational setting: A social network perspective," *Performance Improvement Quarterly*, vol. 21, no. 4, pp. 5–33, 2009.

[24] E. Drymonas, K. Zervanou, and E. G. Petrakis, "Unsupervised ontology acquisition from plain texts: the ontogain system," pp. 277–287, 2010.

[25] G. Wohlgenannt, "Leveraging and balancing heterogeneous sources of evidence in ontology learning," pp. 54–68, 2015.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[27] J. A. Minarro-Giménez, O. Marin-Alonso, and M. Samwald, "Exploring the application of deep learning techniques on medical text corpora." *Studies in health technology and informatics*, vol. 205, pp. 584–588, 2014.

[28] R. Ismail, Z. A. Bakar, and N. A. Rahman, "Extracting knowledge from english translated quran using nlp pattern," *Jurnal Teknologi*, vol. 77, no. 19, 2015.

[29] A. Panchenko, S. Faralli, E. Ruppert, S. Remus, H. Naets, C. Fairon, S. P. Ponzetto, and C. Biemann, "Taxi at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, pp. 1320–1327.

[30] T. Atapattu, K. Falkner, and N. Falkner, "A comprehensive text analysis of lecture slides to generate concept maps," *Computers & Education*, vol. 115, pp. 96–113, 2017.

[31] S. Sbissi, M. Mahfoudh, and S. Gattoufi, "Mapping clinical practice guidelines to SWRL rules," pp. 283–292, 2019.

[32] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," in *Soviet physics doklady*, vol. 10, 1966, pp. 707–710.

OWA Operator for Missing Data Imputation

1st Nassima Ben Hariz
Research Team SIIVA, LIMTIC Laboratory
Institut Supérieur d'Informatique,
Université de Tunis El Manar,
Ariana, Tunisia
nassima.benhariz@gmail.com

2nd Hela Khoufi
Faculty of Computing and
Information Technology
Jeddah University
Jeddah, KSA
hela.khoufi@gmail.com

3rd Ezzeddine Zagrouba
Research Team SIIVA, LIMTIC Laboratory
Institut Supérieur d'Informatique,
Université de Tunis El Manar,
Ariana, Tunisia
e.zagrouba@gmail.com

***Abstract*—Among the most relevant problems effecting the quality of data is the presence of missing data. A panoply of methods, addressing this problem, is proposed in the literature handling different types of data. According to many studies, these methods perform differently and it is preferable to combine them in order to improve their results. Thus, choosing the appropriate combination method to handle missing data is considered as one of the most challenges confronting the researches. In this work, we are interested in handling continuous data by using a novel solution based on the Ordered Weighted Average (OWA) operator to aggregate the results of three imputation methods such as KNN, MissForest and EM algorithm. The choice of the OWA operator is argued by the fact that this operator has shown good results in different fields such as ontology matching and Breast tumor classification. Experiments conducted in this work, using real datasets, show that OWA achieves good results for imputing missing data.**

***Index Terms*—Missing Data, Combination, OWA operator, Imputation Methods, KNN, MissForest, EM Algorithm.**

I. Introduction

In real-world problems, data are generally characterized by their imperfection. One of the most common forms of imperfection is missing data. Different sources can be the origin of the existence of missing data in databases citing for instance the death of patients in medical domain, equipment malfunctions in industrial field, refusal of respondents to answer certain questions in surveys, loss of files, etc. This problem has gained a growing interest in recent researches in data mining trying to reduce the risk of incorrect estimates and results.

In fact, for handling missing data, many techniques are proposed in the literature [1], [4], [14]–[16], [18]. The simplest solution is the deletion of any observation with missing values. This method can be applied only with few number of missing data. Another solution consists in imputation strategy that replaces each missing case with a plausible value (single imputation) or with a vector of plausible values (multiple imputation [14]). For single imputation, the most popular methods are (1) the mean imputation [9] which replaces missing values of a variable with the mean of all known values of that variable (2) the regression imputation [9] which uses the parameters of the regression model on the available values to estimate missing values and finally (3) the K-Nearest Neighbors (KNN) imputation which replaces the missing values with the value of K nearest neighbors in the data [4], [15]. Another category includes iterative methods based on iterative algorithms such as MissForest [18] and Expectation Maximization (EM) algorithm [6].

In machine learning domain, it is demonstrated that the combination of algorithms can provide a better performance than the use of each algorithm separately [17]. In this scope, we have studied the effect of the combination in the imputation process and we have proved experimentally that the weighted mean combination of MissForest with other imputation methods give the best results [3]. In this paper, we studied another combination method based on the Ordered Weighted Average which proved its efficiency in different fields such as data mining [19], neural networks [5], ontology matching [8] and breast tumor classification [11]. We discuss the conditions under which the OWA combination is most effective comparing to the individual results and we compare its results with the results obtained by weighted mean combination and single MissForest.

We are particularly interested in the combination of three methods which are KNN, MissForest, and EM algorithm considered among the most efficient methods for the imputation problem [13].

This work is organized as follows: Section II describes the imputation problem. Section III gives a short presentation of the OWA operator. Section IV explains our proposed approach based on the OWA combination of imputation methods. Section V presents a comparative study using different real datasets. Finally, the concluding discussion and remarks are provided in Section VI.

II. Imputation problem

Assume that $X = (X_1, X_2, ..., X_p)$ is a $n \times p$ dimensional database with n rows representing the observations $(i_1, ..., i_n)$ and p columns representing the variables $(v_1, v_2, ..., v_p)$. In a database, missing values are modeled by empty cases or by other notations such as '?' or 'NA'. If X_{comp} is a matrix of complete data, we can artificially introduce missing data for X_{comp} and we obtain a matrix with missing values denoted by X_{mis}. Imputation methods for missing values replace missing cases x_{mis} with a plausible values to obtain an imputed matrix denoted by X_{imp}. Figure 1 shows an example of X_{mis} dataset extracted from the iris database [1] with eight missing values.

[1]the description of this dataset will be detailed in section V

X_1	X_2	X_3	X_4	Class
4,6	3,4	1,4	0,2	NA
5	NA	1,5	0,4	setosa
NA	3,1	1,5	0,1	setosa
5,4	3,7	1,5	0,2	setosa
4,4	2,9	1,4	NA	setosa
NA	3,1	1,5	0,1	setosa
6,5	2,8	4,6	NA	versicolor
NA	2,8	4,5	1,4	versicolor
6,7	2,5	5,8	1,8	virginica
7,2	3,6	6,1	NA	Virginica

Fig. 1: Example of observations with missing values extracted from iris dataset

X_1	X_2	X_3	X_4	Class
4,6	3,4	1,4	0,2	NA
5	NA	1,5	0,4	setosa
NA	3,1	1,5	0,1	setosa
5,4	3,7	1,5	0,2	setosa
4,4	2,9	1,4	NA	setosa
NA	3,1	1,5	0,1	setosa
6,5	2,8	4,6	NA	versicolor
NA	2,8	4,5	NA	versicolor
6,7	2,5	5,8	1,8	virginica
7,2	3,6	6,1	2,1	Virginica

Legend:

x^4_{obs}

x^4_{mis}

y^4_{obs}

y^4_{mis}

Fig. 2: Different parts of the example extracted from iris dataset for the variable X_4

For an arbitrary variable $X_s(s = 1, ..., p)$ from X with missing values at entries $i^s_{mis} \in \{1, ..., n\}$ the dataset can be divided into four parts:

1) The observed values of X_s, denoted y^s_{obs};
2) The missing values of X_s, denoted y^s_{mis} ;
3) The variables other than X_s with observations $i^s_{obs} \in \{1, ..., n\} \backslash i^s_{mis}$, denoted x^s_{obs} ;
4) The variables other than X_s with observations i^s_{mis} , denoted x^s_{mis}.

Figure 2 presents the four parts described above of the same example from iris database for the variable X_4 (with $s = 4$).

III. Combination Operator OWA

The Ordered Weighted Average (OWA) is considered as powerful aggregation operator used to combine multiple inputs coming from different sources of information. This operator is introduced by Yager in [20] and used in many application fields such as data mining [19], fuzzy logic controllers [21], neural networks [5], ontology matching [8] and Breast tumor classification [11].

A. Definition

Formally, an OWA operator is mapping from R_m to R, $R = [0, 1]$ defined by :

$$F(a_1, a_2, ..., a_m) = \sum_{i=1}^{m} w_i b_i \quad (1)$$

where

- $(a_1, a_2, ..., a_m)$ is a set of m arguments, $a_i \in [0, 1], 1 \leq i \leq m$.
- $b_i \in (b_1, b_2, ..., b_m)$, the set of arguments obtained by ordering $(a_1, a_2, ..., a_m)$ in a descending order.
- $w_i \in (w_1, w_2, ..., w_m)$ the weights of the OWA operator, $w_i \in [0, 1]$, $\sum_{i=1}^{m} w_i = 1$.

It is important to notice that the operators maximum, minimum and mean are three special cases of OWA operator:

- $F(a_1, a_2, ..., a_m) = max(a_1, a_2, ..., a_m)$ if $w_1 = 1$ and $w_j = 0$ for $j \neq 1$.
- $F(a_1, a_2, ..., a_m) = min(a_1, a_2, ..., a_m)$ if $w_m = 1$ and $w_j = 0$ for $j \neq m$.
- $F(a_1, a_2, ..., a_m) = mean(a_1, a_2, ..., a_m)$ if $w_j = \frac{1}{m} \forall j \in [1, m]$.

The use of OWA operator is generally composed of the following three steps [20]:

1) Reordering the input arguments in descending order.
2) Determining the weights associated with the OWA operator by using a proper method as the linguistic quantifiers.
3) Using the OWA weights to aggregate these reordered arguments.

B. Weight determination

The weight w_i, $i \in [1, m]$ is associated with a particular ordered position i of the arguments. To determine this weight, we used the linguistic quantifiers developed by Yager [20], since these quantifiers gives semantics to the different weights. The weights are computed as follows:

$$w_i = Q(i/m) - Q((i-1)/m), i = 1, 2, ..., m \quad (2)$$

where

- m is the number of arguments.
- Q is the nondecreasing proportional fuzzy linguistic quantifier defined as the following:

$$Q(r) = \begin{cases} 0, & if r < a; \\ (r-a)/(b-a), & if a \leq r \leq b; \\ 1, & if r > b, \end{cases} \quad (3)$$

where $a, b, r \in [0, 1]$, a and b are the predefined thresholds.

IV. Proposed approach

We are interested in the combination of the three methods KNN, missForest and EM algorithm because they use continuous data. In the following, we present a brief description of the principle of these methods. Then, we explain our approach based the OWA combination of these imputation methods.

A. Used Methods

1) KNN Imputation.: Thanks to its simplicity, KNN imputation is the most well-known method for continuous datasets [4]. The goal of this method is to replace the missing values with the value of k nearest neighbors in the dataset. For each observation x_i with missing values, k nearest values are selected by calculating the distances between x_i and the other observations $x_j (i, j = 1, ..., n, j \neq i)$. The most used distance measure is the Euclidean distance, that is calculated by the following equation:

$$EuclideanDistance(x_j, x_i) = \sqrt{\sum_{s=1}^{p} (x_j^s - x_i^s)^2} \quad (4)$$

For continuous variables, imputed value is the weighted average of the k neighbor values. In the case of discrete variables, the missing value is replaced by the most frequent value among the values of the K nearest neighbors. This method preserves the distribution and the correlation between variables.

2) MissForest.: Proposed by Stekhoven and Bühlmann [18], MissForest (MF) is based on the algorithm of Random Forest (RF) [2]. Initially, MF estimate missing values using mean imputation or any other imputation method. Then, sort all the variables $X_s (s = 1, ..., p)$ starting with the variables having the lowest rate of missing values. For each variable X_s, pick a RF with response y_{obs}^s and predictors x_{obs}^s and estimate the missing values y_{mis}^s by applying the trained RF to x_{mis}^s (see section 2). The process continue until a stopping criterion ν is reached (see algorithm 1). Indeed, the stopping criterion ν is obtained when the difference between the newly imputed data matrix X_{imp}^{new} and the previous one X_{imp}^{old} increases for the first time. For continuous variables N, this difference is defined as:

$$\Delta_N = \frac{\sum_{j \in N} \left(X_{imp}^{new} - X_{imp}^{old} \right)^2}{\sum_{j \in N} \left(X_{imp}^{new} \right)^2} \quad (5)$$

The performance of the method was assessed using Normalized Root Mean Squared Error (NRMSE) proposed by [12]. For continuous variables this measure is defined by the Equation 3 where μ, σ^2 are the empirical mean and the variance computed over the continuous missing values.

$$NRMSE = \sqrt{\frac{\mu \left((X_{comp} - X_{imp})^2 \right)}{\sigma^2 \left(X_{comp} \right)}} \quad (6)$$

3) EM Algorithm.: The EM algorithm is a probabilistic approach to estimate missing values [6]. It is an iterative method for maximizing the observed likelihood. The principle of this method is illustrated in the algorithm 2. Indeed, this method has two steps :

1) Expectation step: computing expected values for the missing data given the current parameter estimates P.
2) Maximization step: estimating new parameter values by maximizing the observed likelihood after substituting the values estimated in step 1.

Algorithm 1 MF Algorithm

```
Inputs: X an n × p matrix and stopping criterion ν
Make initial estimation for missing values;
k ← vector of sorted indices of columns in X;
while not ν do
    X_imp^old ← store previously imputed matrix
    for s in k do
        Pick a random forest using y_obs^s and x_obs^s
        Estimate y_mis^s using x_mis^s
        X_imp^new ← update imputed matrix using estimated y_mis^s
    end for
    update ν.
end while
return the imputed matrix X_imp^new
```

Algorithm 2 EM Algorithm

```
Random initialization for parameters P
while the algorithm not converge do
    Estimation of X_mis according to X_obs and the values of P
    Maximizing the likelihood using the estimated X_mis performed in the previous step.
    Updating the values of P
end while
```

B. Combination approach

The combination of classifiers is a classical approach that has proved its efficiency in several problems on machine learning and pattern recognition researchers [17]. The main idea of this approach is to combine the results of two or more classifiers to obtain a single and relevant result.

The most commonly used combination methods are: weighted mean, minimum, maximum and product for numerical data, and majority vote for categorical data [17]. In [3], we experimentally proved that the weighted mean combination of MissForest with another ensemble of methods such as EM or EM and KNN gave the best results.

Considering the importance of the OWA in different fields such as ontology matching [8] and Breast tumor classification [11], we propose to combine the outputs of the imputation methods using this operator.

The main steps of the combination approach using the OWA method are presented in algorithm 3. We start by normalizing the imputed datasets by the different methods in a range between 0 and 1, since OWA is applied for values in $[0, 1]$. Then, we store these normalized datasets in a same dataframe containing vectors formed with m values obtained from m results of imputed methods. For each vector from the dataframe, we calculate the weights of each value using

the method cited in Section III-B. After this, we calculate the ordered weighted average for each vector using Equation (1) to obtain a new imputed dataset X_{comb} with the OWA combined values. Finally, we denormalize this dataset to be able to compare it with the original dataset by computing the error rate.

Algorithm 3 Combination Algorithm

Inputs: X_{imp} : m $(n \times p)$ imputed databases by different methods
Normalize X_{imp}^{k}, $k \in [1, m]$
for i in 1 to n **do**
  for j in 1 to p **do**
    $V \leftarrow (X_{imp}^{1}[i,j], X_{imp}^{2}[i,j], ..., X_{imp}^{m}[i,j])$
    Calculate $W = (w_1, w_2, ..., w_m)$ using Equation (2)
    Calculate $OWA(V, W)$ using Equation (1)
    $X_{comb}[i,j] \leftarrow OWA(V, W)$
  end for
end for
Denormalize X_{comb}
Return X_{comb}

V. Experimental Results

The main objective of the experiments conducted in this work is to evaluate the performance of the OWA operator. We compare this method with KNN, MissForest, EM Algorithm and weighed mean combination. The experiments were performed using six continuous datasets extracted from UCI (University of California Irvine) [10].

A. Database description

Iris dataset is a very popular dataset introduced by Fisher [7] to illustrate linear discriminance analysis. Each row of the dataset represents an iris flower, including its species (setosa, versicolor, and virginica), the length and the width of its botanical parts, sepal and petal, in centimeters.

Glass identification dataset is a classification dataset. The study of classification of glass types was motivated by criminological investigation. This dataset contains the description of 214 fragments of glass representing examples of the chemical analysis of 7 different types of glass.

Seeds dataset is introduced to illustrate measurements of geometrical properties of kernels belonging to three different varieties of wheat: Kama, Rosa and Canadian. For each element in the dataset, seven geometric parameters of wheat kernels were measured.

The Vertebral Column dataset is a Biomedical dataset having six biomechanical features used to classify orthopaedic patients into three classes (normal, disk hernia or spondilolysthesis) or two classes (normal or abnormal).

The Yeast database contains information about a set of yeast cells. Its objective is to determine the localization site of proteins in a yeast's cell using various descriptors. Each protein has to be classified into one of ten different cellular components.

For Banknote dataset, the examples were extracted from banknotes images. These images were taken of 1372 banknotes. Wavelet transformation tools were used to extract the descriptive features of the images (Variance, Skewness, Kurtosis, and Entropy) to classify banknotes into two classes (genuine and not genuine).

Table 1 summarizes the properties of the selected datasets.

These datasets have no missing values. This means that we have access to a complete matrix X_{comp}. The main reason for this choice is that we want to have total control over the missing data in the dataset. For instance, we would like that the test sets do not have any missing data.

In our experiments, missing values were artificially generated with different rates and attributes. From each original dataset, we derived five incomplete datasets, removing 10%, 20%, 30%, 40% and 50% of values completely at random. For each rate, we perform 100 independent runs by randomly generating missingness patterns (see Figure 3). We imputed the missing datasets using the three imputation methods, then we combine the obtained results by the weighted mean and the OWA operator.

To assess the performance of each imputation approach, we calculated the Normalized Root Mean Squared Error, by comparing the original dataset with the imputed datasets (Equation 6). The final NRMSE is averaged over 100 repetitions. Lower values of NRMSE indicate better estimates of the variables. The runtime of the different imputation methods is also compared.

B. Evaluation of imputation methods

The imputation errors of different methods are given in Figures 4, 5, 6, 7, 8 and 9. Overall, our analysis showed that bias was lower when the rate of missing values was lower.Thus, the performance decreased with increasing percentage of missing values in all datasets. We can also see that MissForest performs better than the other two methods in all datasets, sometimes reducing the average NRMSE by up to 20%. For glass dataset (Figure 5), EM algorithm has a slightly smaller NRMSE than MissForest only when the rate of missing values is lower than 15%. But for vertebral column dataset, yeast dataset and banknote dataset, EM performed less well due to the increase of number of instances.

C. Evaluation of OWA operator

In [3] we have proved that the weighted mean combination of MissForest with another methods such as KNN, EM or EM and KNN gave better results than using MissForest individually for continuous data. In this paper, we propose to combine the outputs of MissForest with the outputs of other methods using the OWA operator.

Using the OWA operator for each dataset, we compute the results obtained by the three possible combinations (i.e MF and KNN, MF and EM, or MF, KNN and EM).

To show the performance of the OWA operator, we compared the results obtained by this method to the results ob-

TABLE I: Properties of the selected datasets.

Databases	Number of instances	Number of attributes (all continuous)	Number of classes
Iris Plants	150	4	3
Glass Identification	214	9	7
Seeds	210	7	3
Vertebral Column	310	5	2
Yeast Database	1484	8	10
Banknote Authentification	1372	4	2

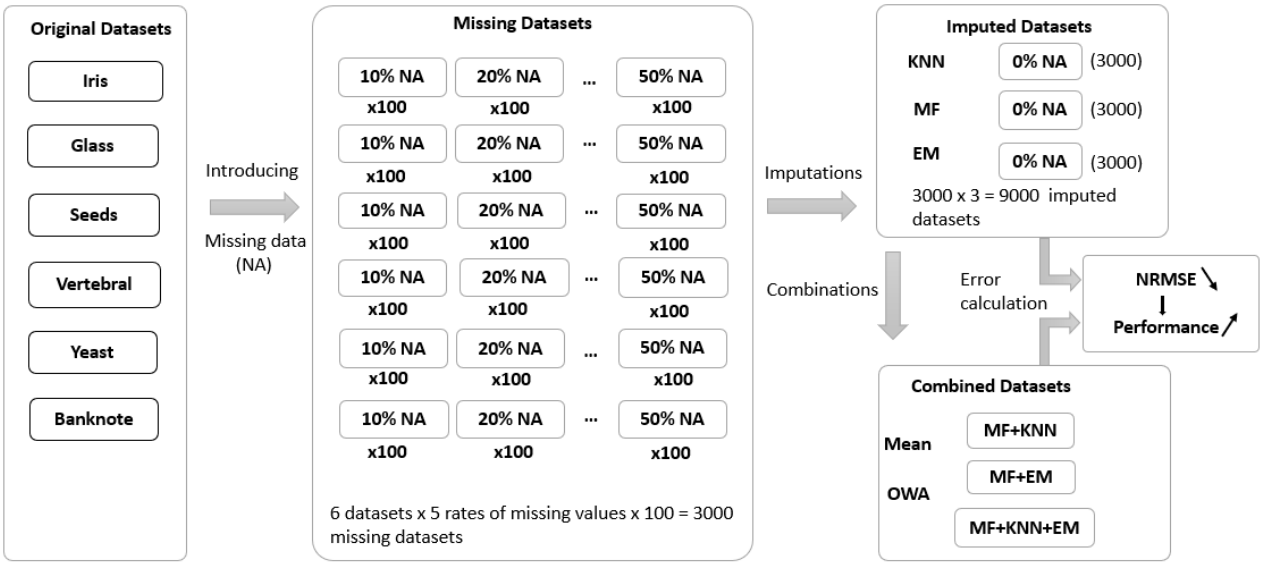


Fig. 3: Main steps used in experimental study.

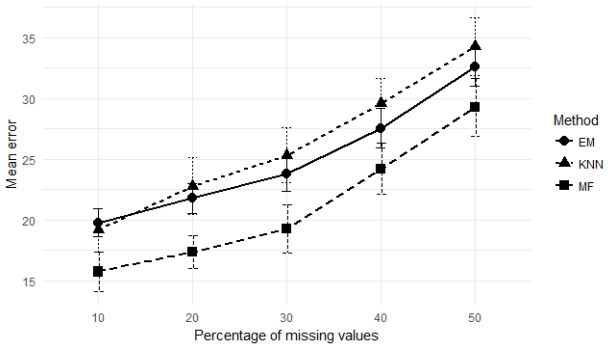


Fig. 4: Results of imputation methods for iris dataset

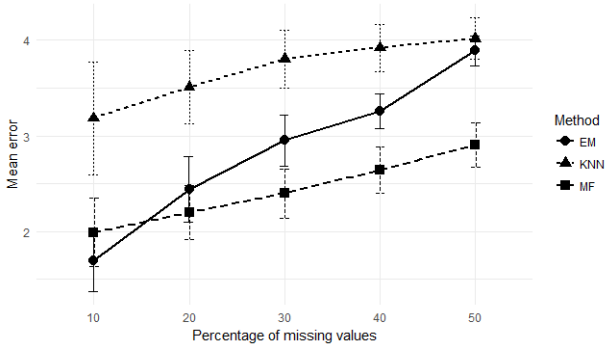


Fig. 5: Results of imputation methods for glass dataset

tained by the weighted mean combination and by MissForest individually (shown previously in [3]).

Figures 10, 11, 12, 13, 14 and 15 present the results of all possible combinations of Missforest with KNN and EM using the mean and the OWA operator comparing with the results of MissForest for each database.

The results of each database are represented by a figure containing three cases. First, the combination of the outputs of MissForest and KNN. Second, the combination of the outputs of MissForest and EM. Finally, the combination of the outputs of MissForest with KNN and EM.

In these figures and in the majority of cases, the worst result is obtained by using a single MissForest.

In Figure 10, we can see that for the three combinations OWA performed well than the mean combination when the rate of missing values is less than 40%. In the two first cases (Fig-

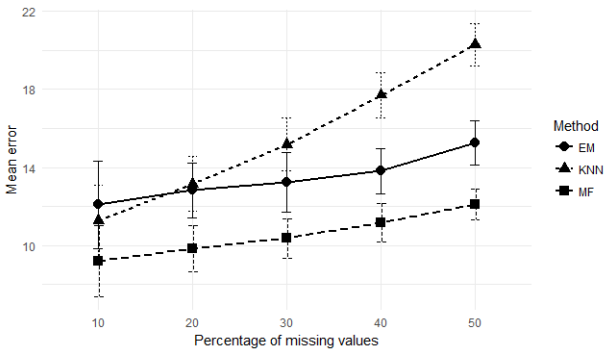


Fig. 6: Results of imputation methods for seeds dataset

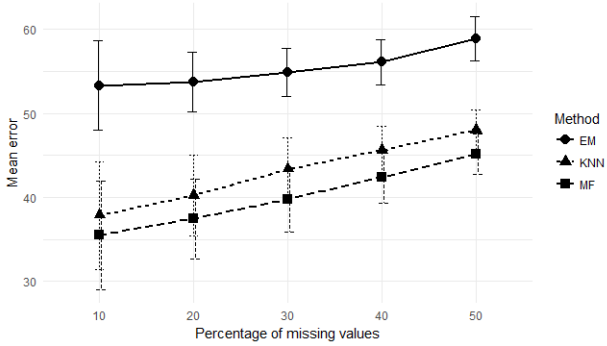


Fig. 7: Results of imputation methods for vertebral column dataset

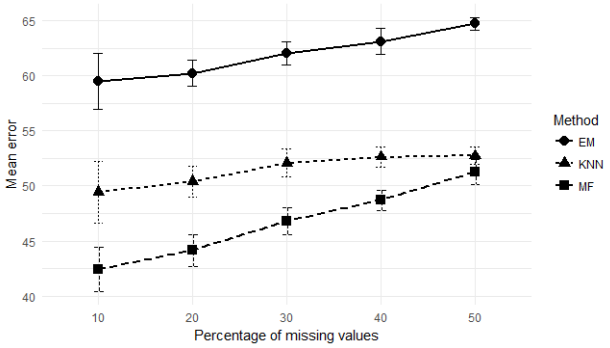


Fig. 8: Results of imputation methods for yeast dataset

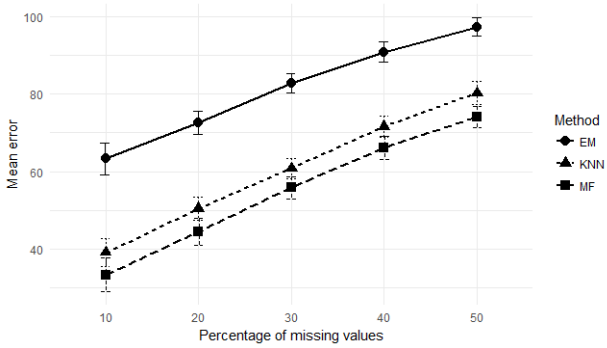


Fig. 9: Results of imputation methods for banknote dataset

ures 10a and 10b), for 50% of missing values, the performance of the OWA combination and the performance of the mean combination are very close, but for the combination of the three methods (Figure 10c) the weighted mean combination provides better results.

For glass database, Figure 11 shows that the performance of OWA is higher than the performance of the mean combination especially for the ensemble composed by the combination of the three methods (Figure 11c).

Also, Figure 12 shows good results of the OWA for the different combinations in seeds database.

For the three other databases (Figures 13, 14 and 15), in the most cases the OWA performed well than the weighted mean combination when the rates of missing values is approximately less than 30%, otherwise the mean combination provides best results.

After the experiments, we remark that to improve imputation results it is preferable to combine MissForest with another imputation method than to use it individually. In this paper, we compared two methods of combination for continuous data: weighted mean and OWA, and we proved that OWA combination is preferable when the rate of missing values is less than 40% especially for small datasets, otherwise we can use the mean combination.

VI. Conclusion

In this work, the behavior of three imputation methods is analyzed according to the rate of missing data into different attributes of six real datasets. The MissForest method provides very good results, even when the training sets have a large amount of missing data. The combination of this method with other methods improves significantly the results obtained by a single MissForest. For low rates of missing values, the combination based on OWA operator is preferable than the weighted mean combination. For high rates of missing data the mean combination can be used.

In the future, it is interesting to study the behavior of the imputation methods for datasets with more instances and with other attribute types as nominal and mixed types (nominal+continuous) and with dependent variables.

References

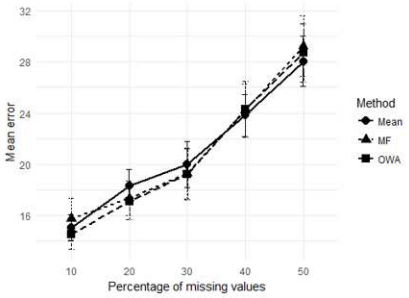
[1] Aljuaid, T., Sasi, S.: Proper imputation techniques for missing values in datasets. In International Conference on Data Science and Engineering (ICDSE). IEEE, 1-5 (2016, August)

[2] Breiman, L.: Random forests. Machine learning, 45(1), 5-32 (2001)

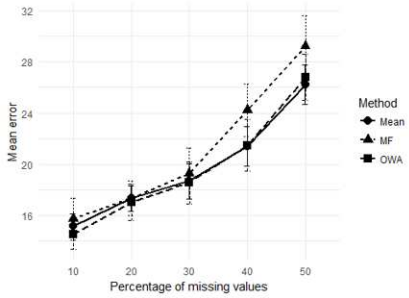
[3] Ben Hariz, N. , Khoufi, H., Zagrouba, E.: On Combining Imputation Methods for Handling Missing Data. In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Cham, 171-181 (2017, June)

[4] Chen, J., Shao, J.: Nearest neighbor imputation for survey data. Journal of official statistics, 16(2), 113 (2000)

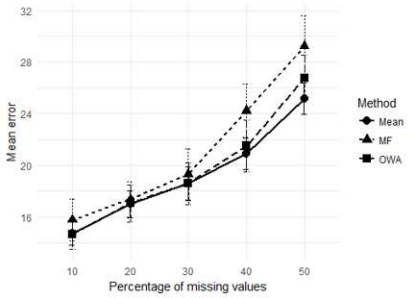
[5] Cho, S. B.: Fuzzy aggregation of modular neural networks with ordered weighted averaging operators. International journal of approximate reasoning, 13(4), 359-375 (1995)



(a) MissForest + KNN

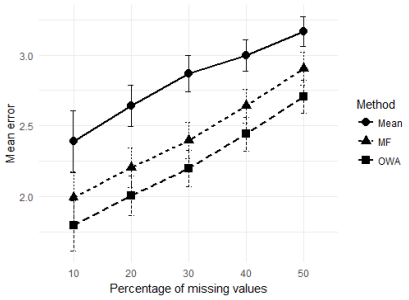


(b) MissForest + EM

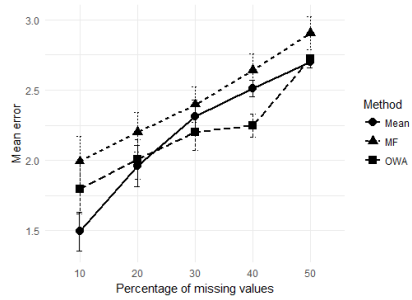


(c) MissForest + KNN + EM

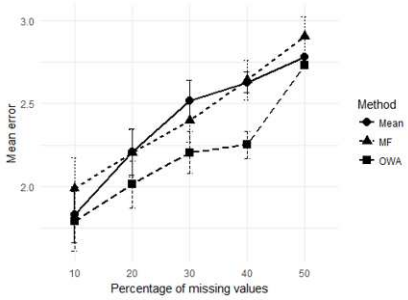
Fig. 10: Combination results for iris database



(a) MissForest + KNN

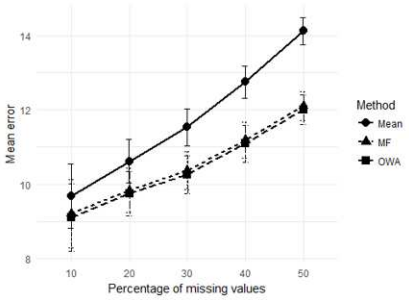


(b) MissForest + EM

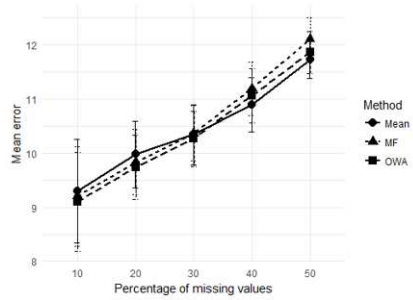


(c) MissForest + KNN + EM

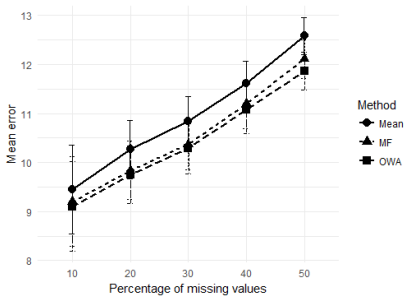
Fig. 11: Combination Results for glass database



(a) MissForest + KNN



(b) MissForest + EM



(c) MissForest + KNN + EM

Fig. 12: Combination results for seeds database

[6] Dempster, A. P., Laird, N. M., Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38 (1977)

[7] Fisher, R. A.: The use of multiple measurements in taxonomic problems. Annals of human genetics, 7(2), 179-188 (1936).

[8] Ji, Q., Haase, P., Qi, G.: Combination of similarity measures in ontology matching using the owa operator. In Recent Developments in the Ordered Weighted Averaging Operators: Theory and Practice (pp. 281-295). Springer Berlin Heidelberg (2011)

[9] Little, R. J. A., Rubin, D. B.: Analysis with missing data. 323-357 (1987)

[10] Merz, C. J., Murphy, P. M.: UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science (1998). http://www.ics.uci.edu/mlearn/MLRepository.html

[11] Mohammed, E. A., Naugler, C. T., Far, B. H.: Breast tumor classification using a new OWA operator. Expert Systems with Applications, 61, 302-313 (2016)

[12] Oba, S., Sato, M. A., Takemasa, I., Monden, M., Matsubara, K. I., Ishii, S.: A Bayesian missing value estimation method for gene expression profile data. Bioinformatics, 19(16), 2088-2096 (2003)

[13] Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., Costa, G. C.: Imputation of missing data in life-history traits datasets: which approach performs the best? Methods in Ecology and Evolution, 5: 961-970 (2014)

[14] Rubin, D. B.: Basic ideas of multiple imputation for nonresponse. Survey Methodology, 12(1), 37-47 (1986)

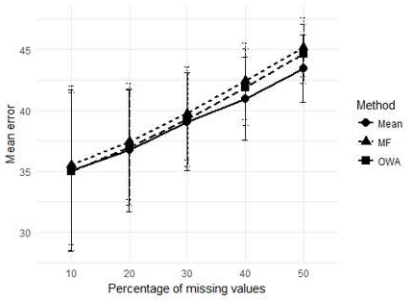
[15] Rubin, D. B., Little, R. J.: Statistical analysis with missing data. Hoboken, NJ: J Wiley & Sons (2002)

[16] Schmitt, P., Mandel, J., Guedj, M.: A comparison of six methods for missing data imputation. Journal of Biometrics & Biostatistics, 6(1), 1 (2015)

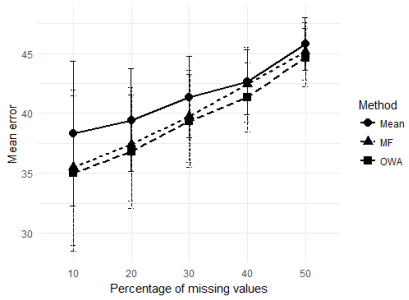
[17] Shipp, C. A., Kuncheva, L. I.: Relationships between combination methods and measures of diversity in combining classifiers. Information fusion, 3(2), 135-148 (2002)

[18] Stekhoven, D. J., Bühlmann, P.: MissForest non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118 (2012)

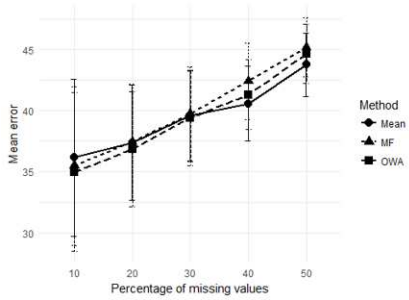
[19] Torra, V.: OWA operators in data modeling and reidentification. IEEE



(a) MissForest + KNN

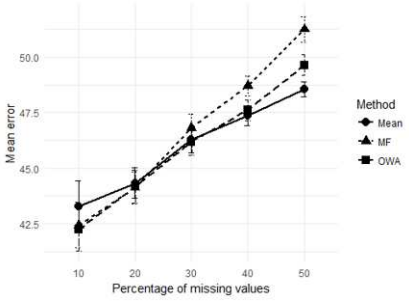


(b) MissForest + EM

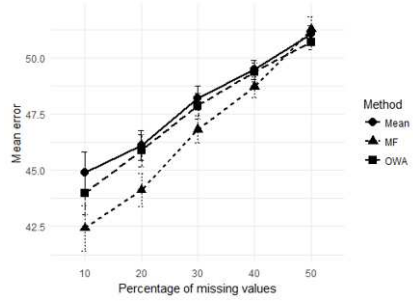


(c) MissForest + KNN + EM

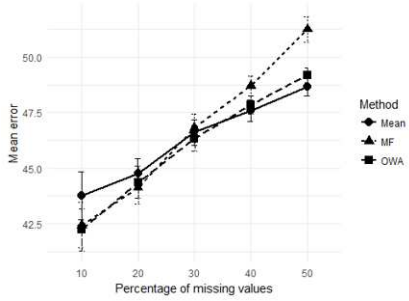
Fig. 13: Combination results for vertebral column database



(a) MissForest + KNN

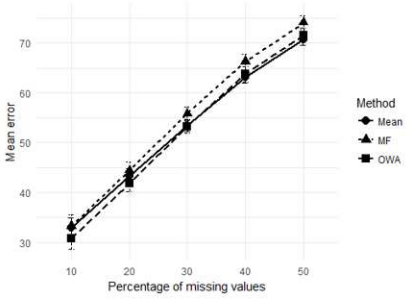


(b) MissForest + EM

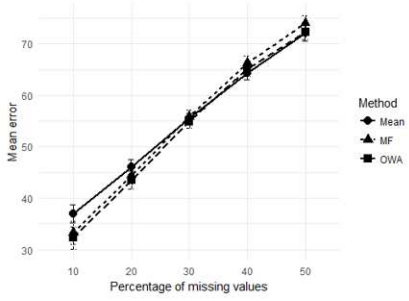


(c) MissForest + KNN + EM

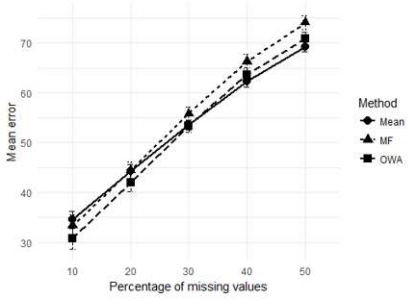
Fig. 14: Combination results for yeast database



(a) MissForest + KNN



(b) MissForest + EM



(c) MissForest + KNN + EM

Fig. 15: Combination results for banknote database

Transactions on Fuzzy Systems, 12(5), 652-660 (2004)

[20] Yager, R. R.: On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Transactions on systems, Man, and Cybernetics, 18(1), 183-190 (1988)

[21] Yager, R. R., Filev, D. P.: Fuzzy logic controllers with flexible structures. In Proceedings Second International Conference on Fuzzy Sets and Neural Networks, 317-320 (1992)

Multi-objective Clustering Algorithm with Parallel Games

Dalila Kessira[1], Mohand-Tahar Kechadi[2]

[1] Laboratory of Medical Informatics, University A. MIRA of Bejaia, Road Targua Ouzemour, Bejaia, Algeria

[2] Insight Center for Data Analytics, University College Dublin, (UCD), Belfield,Dublin 4, Ireland

Dalila.kessira@gmail.com

Abstract

Data mining and knowledge discovery are two important growing research fields in the last few decades due to abundance of data collected from various sources. In fact, the exponentially growing volumes of generated data urge the development of several mining techniques to feed the needs for automatically derived knowledge. Clustering analysis (finding similar groups of data) is a well-established and a widely used approach in data mining and knowledge discovery. In this paper, we introduce a clustering technique that uses game theory models to tackle multi-objective application problems. The main idea is to exploit specific type of simultaneous move games [1], called congestion games with player-specific payoff functions [2] [3]. Congestion games offer numerous advantages ranging from being succinctly represented to possessing a Nash equilibrium [4] that is reachable in a polynomial-time. The proposed algorithm has three main steps: 1) it starts by identifying the initial players (or the cluster-heads); 2) then, it establishes the initial clusters' composition by constructing the game to play and try to find the equilibrium of the game. The third step consists of merging close clusters to obtain the final clusters. The experiment results show that the proposed clustering approach is very promising and can deal with many limitations that other techniques have, such as scalability, performance, and data heterogeneity.

Key words: Data mining, data analysis, clustering, game theory, simultaneous-move games, congestion games with player-specific payoff functions, Nash equilibrium.

References:

[1] D. Parkes and S. Seuken, "Game Theory I : Simultaneous-Move Games," 2013, pp. 17–41.

[2] R. W. Rosenthal, "A class of games possessing pure-strategy Nash equilibria," *Int. J. Game Theory*, vol. 2, no. 1, pp. 65–67, 1973.

[3] I. Milchtaich, "Congestion Games with Player-Specific Payoff Functions," *Games Econ. Behav.*, vol. 13, no. 1, pp. 111–124, 1996.

[4] J. Nash, "NON-COOPERATIVE GAMES," *Ann. Math. Second Ser. Ann. Math.*, vol. 54, no. 2, pp. 286–295, 1951.

Design of a Breast Image Data Warehouse Framework

Nedra Amara
Institute of Management
University of Tunis
Bardo, Tunisia
Email: amaranedra293@gmail.com

Olfa Lamouchi
Engineers National School
of Carthage
45 Rue des entrepreneurs 2035, Tunisia
Email: olamouchi.enicar@gmail.com

Said Gattoufi
Institute of Management
University of Tunis
Bardo, Tunisia
Email: algattoufi@yahoo.com

***Abstract*—A medical image data warehouse framework is needed, which includes medical imaging and textual reports in assisting diagnosis management and research. We propose such a user framework, illustrate general design and system architecture, and describe a breast- imaging data warehouse implemented for breast cancer research. The image data warehouse for breast cancer disease is constructed on top of a picture archiving and communication system (PACS) and relates a star schema for modelisation and recognized structured and unstructured data interface and design norms. The implementation is based on ORB (Object Request Broker) and web technology that separates the source layer, integration layer, data warehouse layer, and graphical doctor interface application. To explain the feasibility of the data warehouse, we propose two evident medical applications-that is to say, clinical diagnostic workup of multimodality breast-imaging cases and research medical data analysis and decision threshold on inflammatory breast cancer lateralization. The medical image data warehouse framework can be adapted and generalized for for some types of cancer disease.**

1. Introduction

Picture Archiving and Communication System (PACS), an obligatory tool in public hospitals, has shown its all-important position in the radiology service for storing and recovering medical images followed by their integration with the radiology information system (RIS) huang2019pacs. The textual integration makes it possible to display on the PACS the radiological reports entered on the RIS attached to the examination of a patient. This transfer to PACS is done automatically once the report is validated in the RIS, either via Health Level 7 (HL7) standardized messages or via the IHE- The retrieve Information for display (RID) profile, if the systems are compatible. The plain text will therefore be accessible directly in the PACS, including remotely, without the possibility of corrections, except to be able to make a copy and a reintegration to PACS in the form of an attached document, functionality that only certain PACS are able to do. It will then be necessary to reintegrate him in the RIS. Reciprocally, some publishers offer a numerical entry of the count directly on the PACS (with or without voice recognition), report which will then also have to be reintegrated into the RIS. In this paper, we describe a breast-imaging data warehouse framework that models and integrates multiple breast-imaging modalities, including digital x-ray mammography, magnetic resonance imaging (MRI), Ultrasound imaging (US), x-ray, computed tomography (CT), [1] and associated clinical and research data to support diagnosis and breast-imaging research. The design of the framework is based on Fast Healthcare Interoperability Resources (FHIR) and recognized industry standards, such as DICOM and Health Level 7 (HL7), in retrieving data from underlying hospital information systems architecture and design. The multimedia data warehouse system supply more powerful use and analysis of multimodality images and records in supporting healthcare hospital research and treatment than what is presently supported by clinical information systems, PACSs, or other medical databases [2].

1.1. Picture Archiving Communication Systems

To meet the challenges of archiving, acquiring, displaying, and communicating medical images, PACSs created during the last decennary as image management systems. First-generation PACSs were created to maintain direct radiologic soft-copy readings but were not destined to include other related textual data sources of radiologic use. Various hundreds of the first-generation systems are actually installed in hospitals worldwide. The ulterior evolution of second-generation PACSs, also named as hospital integrated PACSs (HI-PACSs), contain the integration of PACSs with heterogeneous information systems like radiologic information system (RIS) and hospital information system (HIS). PACSs have their place in the enterprise to offer more cost-effective and efficient access to all medical information to assist patient care [3].

1.2. Expanding medical Imaging Services on the outside Radiology interpret

Scientist and medical staff integrated into other departments and outside radiology departments need to have as simple

access to the medical images archived in PACSs as do radiologists themselves. Nowadays use PACS demonstration workstations have no viability for warehousing textual diagnostic reports [4].
These workstations are also strongly employed by radiologists and are hard to enter by scientists and medical staff who are not managing dealy diagnostic output. While the working of PACSs has led to the buildup of large stores of multimedia data, containing diagnostic reports, medical images, and patient data, analyzing and obtaining these data outside the scope of quotidian hospital operation is a tedious and long procedure. What is the actual application of analyzing medical images for diagnosis and scientific research? Researchers and scientists use robust computer workstations to recuperate digital images from a PACS and other medical-imaging sources, stock images on local disks, display multiple image sets at the same time, and use image-processing routines such as quantitation, registration, and segmentation. The medical data are inefficiently managed through level files on a hard disk, and there is no centralized to assure data coherence [5]. The analysis task must be accomplished on a unique machine; Such obstacles prevent the capacity of scientists outside a radiology service to collect and examine images when preparing a diagnostic examination of patients for tumor surgery and defining decision thresholds. in this work, we will focus on data centralizing, the PACS does not allow data centralizing [6] [7]. Doctors pass an important deal of time, frequently weeks, collecting data from various information systems and paper records before sitting down to do data analysis or diagnostic examination. Thanks to data centralization, the medical staff can access any services, they always have all the data to do the medical analysis.The Centralization avoids data loss. Thus, To solve this problem, we will propose a data warehouse [8]. Our medical image data warehouse are proposed to remedy the data decentralizing problem and facilitate diagnostic processes [9].

2. The state of the art

The manuscript review the state of the art of research data warehouse for medical imaging. Many tools and frameworks already exist, often already implementing some of the needs the authors identify. The Extensible Neuroimaging Archive Toolkit (XNAT) is an archiving software platform conceived to ease mutual management and processing tasks for neuroimaging and related data, supplying secure warehousing and access layer. XNAT's architecture keeps a three-tier conception pattern that integrates a relational database backend, a web-based user interface, and a Java-based middleware engine. [10] Today, there are various publicly disposable solutions to manage clinical and medicals data more efficiently than XNAT, such as integrating data for analysis, anonymization, and sharing (iDASH) is a cloud-based platform for sharing and development of tools and algorithms and for secure HIPAA compliant data sharing [11] [12],The Cancer Translational Research Information Platform (caTRIP) instigates medical aggregation, linking common data elements [13], and meta-data navigation, TCIA-QIN ease medical data sharing of multi-site and medical imaging collections and complex clinical data, the integrated clinical omics database (iCOD) are based on environmental, clinical, and pathological data obtained from patients who admitted medical care [14], cBio Cancer Portal is open-access resource permitting an interactive search of multidimensional medical data sets [15], Biology-Related Information Storage Kit (BRISK) is a set of several web-based data management tools that supply cohesive data integration and management platform. It was specifically conceived to supply the architecture necessary to instigate collaboration and accelerate data sharing between scientists [16], A Systems Medicine Platform for Personalized Oncology (G-DOC), and tranSMART enable all type of users to collaborate, use the best analytical tools, communicate and establish convergent standards [**?**], and instigate new informatics-allowed translational science in the academic, pharmaceutical, and not-for-profit sectors [17] [10], existing the choice to design a data warehouse system and leave XNAT in charge of neuroimaging data. Preface A growing number of international projects are generating a big amount of neuroimaging and related data such as behavioural, genetic or clinical database and management systems. The period researchers pass to manage and query the data is growing with the complexity and size of these databases. To automate medical data management and treatment tasks, it is focal to be able to make access to a database. [18] propose a PyXNAT, a Python module that interacts with (XNAT) through Python appeal across various operating systems. [19] proposes a data warehouse evolution framework. This framework was created by increasing the data warehouse adaptation framework [19], which was previously conceived. The framework could automatically detect changes in the model of data sources and adapt a data warehouse model and ETL processes, according to the admin decision [20]. in this work, [21] are proposed the realization of a data-warehousing framework(HMAF), in order to help health administrators of the public healthcare system supplying data and tools containing technique to perform involved analysis. the authors can achieve that the objective of the study in implementing a framework to sustain analysis and enhance the decision-making tasks was accomplished in an efficient and satisfactory way, since HMAF possible grade data integration and ease and makes as flexible as the information analysis process, on the health management environment. This framework demands amelioration regarding data inserts and updates in the Health Map Operational Data Store (HMODS). The supplying of medical data by the cities is essential for the data maintenance of HMAF. With updated data on HMODS, consequently, when a new charge is performed in HMDW, it is possible to make analysis on current data as well as on historical data. In This work [22] represents a new framework, named the integrated Genomics Anesthesia System (iGAS), to exploit and integrate the richness of clinical data collected during the perioperative time with genomic data from the member in the Icahn School of Medicine at Mount Sinai's biobank. The authors [23]

have designated an entire, multi-tier image data warehouse framework and the matching software development process for relevant neuroimaging research and large-scale medical data analysis. This medical framework is founded on the object-oriented analysis and design (OOAD) methodology. this information framework can be generalized for new medical imaging domains. For their prime implementation, the focus is clinical epilepsy. the researchers are steadily reaching the scope of the application to integrate data of other disease types, such as digital mammography, brain tumors, and lung nodule cancer. [24] have implied a slight variation of the dimensional modeling technique to make a data warehouse more suitable for healthcare applications. One of the essential aspects of conceiving a healthcare data warehouse is finding the appropriate grain for various levels of analysis. the authors have proposed three levels of grain that permit the analysis of healthcare results from highly outline reports on part of care to fine-grained studies of advance from one care visit to the next. These grains permit the database to maintain multiple levels of analysis, which is crucial for healthcare decision making. Healthcare processes that imply multiple health testing measures across several patient treatments make result analysis extremely difficult.To resolve this problem, [2] realize a lean variation of the dimensional modeling technics to develop a data warehouse more suitable for healthcare outcome research. To resolve this problem, the conception process imply of the following steps: specify and determine the healthcare process to model, determine the result grains of the healthcare process and specify data marts and dimensions to catch the healthcare process, specify dimensions and their hierarchies, specify the measures and fact table, and realize the design for a special On-Line Analytical Processing (OLAP) system. figure1 illustrate some sharing data platforms.

3. Dimensional modeling

This informatics idea use methods and norms to the design of a medical image data warehouse framework for data analysis, multimedia operation, data analysis, and research. The framework is an absolute representation of the problem that summary the numerous aspects of breast cancer abnormality tumor treatment and research to a high stage of visual comprehension. It can be revised and details to supply new utilities or maintain a new application field. The design model and modular architecture employed in the framework will be helpful in resolving problems in another's medicals domain.

3.1. Medical imaging data warehouse system

Currently, the elementary intent of most database systems was to face the needs of operational database systems, which are generally transactional in nature. Nowadays, Hospitals collect and generate and a vast quantity of data in the continuous procedure of providing treatment [25]. Traditional patterns of operational systems in hospitals contain image presentation and reading, image reports and scheduling, patient record and order entry. Operational systems are concerned essentially with the processing of a unique transaction and are optimized for transactional updates. Standardization is a term for a handle the data modeler goes through to avert the traps simply meet when attributes, the connection of entities, and relationships that make up a relational data model that is regularly employed in today's database systems. Usually, the needs of an operational system do not change a lot.

On the other hand, a data warehouse implementation is different, as a conventional transaction process with big quantity of data, which are aggregated in nature. The operational process are frequently de-normalized in the data warehouse, a process in which the construction of the tables from the operational environment is changed to permit quicker treatment in the data warehouse [26]. The data stocked in the data warehouse undergo little if any update action. Their principal goal is to be read by users although they make decisions or check the hypothesis.

Under today's competitive care domain, health care decision-makers must be able to interpret movements, research factors; use data based on clear and research factors, timely data displayed in a meaningful format. Researchers would utilize the data warehouse to check scientific hypotheses and treat with "What if?" questions in considering many aspects of disease management and care.

3.1.1. The primary components of multimedia data warehousing solutions. The data warehousing procedures include five-phase—get ready the data, load and obtain the data into a data staging area, extract and transform the data, records the data into the data warehouse, analyze the data, and reporting the data{8745864. The principal tasks in each processing element are described in figure 2. Many characteristics differentiate medical image data warehouse from the data warehouse used in other business industries or company. A medical imaging data warehouse would be created more handily in a PACS or numerical imaging service environment in which a large mass of images and related textual reports are obtained and stocked centrally [27]. The accessibility of message exchange norms and interfaces such as HL7 and DICOM also aid in diminishing the complexity of data preparation and acquisition [2].

Figure 2 illustrates the step flow for designing medical image data warehouses. The image data warehousing development repose of five parts—make the data, receive and load the data into a data staging area, transform and extract the data, store the data into the medical data warehouse servers, and analyze and mine the data. The crucial tasks in each processing part are designated in the figure. Make the Data: Determine the study objectives or hypotheses. Select the target data, patient population, databases, and study protocols.

Receive the Data: receive and charge medical images and data archives from various hospital systems into a data staging area. Scan paper documents. it any.

Transform and Extract the Data: Cleanse the data Transform the data into formal appropriate for data treatment and

Authors	Description	Benefits	Limits
XNAT [10], [11]	✓ XNAT was conceived to capture data from multiple sources, to keep the data in a secure medical repository, and to diffuse the data to permitted users	✓ Aid common productivity and management tasks for neuro-imaging and related data. ✓ Secure Access to user interface and data archive ✓ Xnat keeps a history profile to follow all modifications made to the managed medical data ✓	❖ It is important to note that the archive is actively managed by domain-knowledgeable staff. ❖ The archive is managed by personnel who know the domain; they are not accessible by everyone. It is difficult to manage the archive by everyone ❖ Data decentralization, Developing technic to permit exchange and communication between systems will be crucial. The potential scripts in which systems will require sharing data resources within databases, and contributing data to a centralized Medical data warehouse.
IDASH [11][12]	Through these several mechanisms, iDASH tools its aims of supplying behavioral and biomedical researchers with access to software, data, and a high-performance computing environment, thus permitting them to test and generate new hypotheses.	✓ Higher privacy and protection of data. All these data require to be secured by privacy-preserving algorithms. ✓ iDASH concentrates on tools and algorithms for sharing data in a privacy-preserving way.	❖ Data compression is obtained by limiting the functionality of the performed computation. ❖ The data is not centralized.
caTRIP [13]	caTRIP enables clinicians to issue through from patients with similar features to get treatments that were managed with success.	✓ Can aid inform care and enhance patient treatment, as well as allow searching for tumor tissue, locating patients for clinical trials. ✓ Understandable by technical and no-technical users such as a clinician	❖ Limited domain, it manages cancer data

Figure 1. Table for some frameworks

archiving medical images. Extract keywords and phrases, and quantitate image features. Add annotations and comments.
Store the Data: Organize and model the data. Check the quality and integrity of the data Archive the data into the data warehouse servers.
Analyze and supply the Data: Analyze the data about the study objectives or hypotheses. Develop end-user Applications, ad hoc query tools, and advanced algorithms to reveal relationships and trends in the data warehouse. Use predictive information to make proactive clinical decisions.
Diverse characteristics differentiate medical image warehousing systems from the data warehousing systems employed in different business company or industries:

- A medical imaging data warehouse implies refined effort of image processing, recording, extraction of image features. It also needs the representing of multimedia data in different forms—contain text, structured reports, zoomable and volumetric and planar medical images, graphics, and scanned paper. The capacity to research all forms of data, quantitatively and qualitatively, not just numbers in records and simple text, generate the image data warehouse an accurate business information door;
- An image data warehouse accent the preparation and obtaining of data with predefined protocols, more than the mining of data or retrospective analysis. The accent is due to the main paradigm adopted in medical research nowadays;
- An image data warehouse offers analytic and statistical mechanics to support a check-based approach in which the user hypothesizes about specific data relationships and then utilize the mechanics to check or disprove the hypotheses. This continues the hypothesis-driven model adopted in most medical research endeavors.

Data mining, in opposition, utilizes what are named discovery-data approaches, in which motif matching and other algorithms are used to define key relationships in the data. The capacity to automatically discover based information invisible in the data and then put it in the accurate format is a critical additional technology to a check based approach. The differences between an operational system and a data warehouse system take place at the very structure by which these technologies are wended. System conception life cycle is the same methodology employed to make any operational system [28] [29].The objective of a data warehouse is to be flexible relatively to trade with the

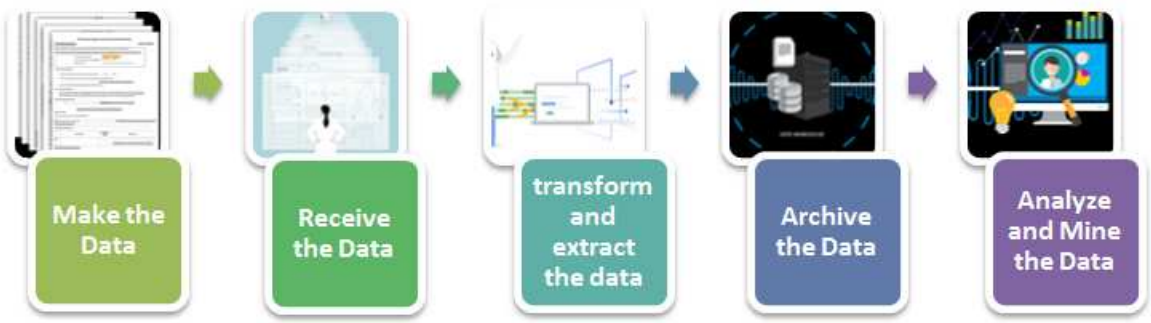


Figure 2. The elementary part of image data warehousing solutions

various needs of the enterprise so that as we posed "what if?" questions, we can go more and more into the data warehouse. Each question asks the underlying information in new ways. If this is done rightly, we will recuperate the responses we seek.

3.2. The star model

The advanced form of medical data warehouses is complicated and time-consuming to survey a set of patient records. Still, they are going to be effective data warehouse presenting to yield a quality patient treatment. The data integration process of the medical data repository is complex scenarios when conceiving medical data warehouse architecture. On this day, any doctors could remember the state of their patients. Still, today, no doctor can proceed with the explosion of health and medical information. Whereas health care institutions have recognized the employ of computers, but in comparison to other domain, its application in healthcare domains has not been encouraging. On account of other factors, it takes much time to get information in various cases; there is no facile accessibility to medical data. But once the medical data warehouse is ready, it merits spending money and the time in it.

With the actual advents, the medical domain associated with the health cycle necessitate higher attention. The big problem allowing of is varied dimensionality, ranging from medical images to numerical form of medical data which needs to respond. founded on the same we propose for an appropriate Medical dimensional model (Figure 3 - logical data model illustration of Medical dimensional model) for the structure of a medical data warehouse which would record the data. The determined model has been conceived using Erwin data modeller version 9.8 [30] and is showed as a star schema . While conceiving the model it has also been taken into regard that the medical data in some dimension may change, for which further attributes have been added, such that it can deed as slowly changing dimension [30]. The purpose of building this data warehouse is to conduct to a platform for using data mining technique to discover correlation among diverse attributes, using combination mining studies, etc. which would assist us in understanding latest translational paradigms which could be employed by doctors, other health professionals and even by a common person who has got knowledge about how to employ computer and internet.

The proposed design includes two fact tables PATIENT_IMAGE_DETAIL and PATIENT, which register the textual (measures) and numerical facts get from the medical images respectively. In the determined dimensional model, the FACT_PATIENT table is referencing to DIM_PATIENT, DIM_DATE, DIM_DISEASE, DIM_DIAGNOSTIC_TEST and Dim_Time. The fact table PATIENT_IMAGE_DETAIL is attributed to DIM_PATIENT, Dim_DATE, DIM_IMAGE, and DIM_TIME respectively. Patient_Id is the primary key of the DIM_PATIENT dimension, which is the unique id that is offered to each patient and this is the id that is joining all other information associated to that patient. The dimension further integrate other descriptive medical information related to a patient like a name, age, education, income, gender,smoking_status, etc. Keeping into regarding the data in the DIM_PATIEN dimension may modify, Start_date, End_date attributes have been inserted so that it can deed as slowly changing dimension. While conceiving the medical dimensional model the historical pursuit prospect was taken into consideration, hence Time and date dimension was included. Date_Id serve as the primary key for the DIM_DATE dimension, which attributes a unique id that is offered to each date. The dimension also integrates diverse attributes like a day of the week, week of the month, calendar year, etc., which can assist to make an analysis in view of the various periods. Time_Id serves as the primary key for Dim_Time which attributes unique id in view of each hour, minute and second. Separate impaction of DIM_TIME dimension ASSURE irrespective of the number of times a test is attended for a patient on any given date, each fact would be stored uniquely in the FACT_PATIENT table. Test_ID and Disease_Id serve as the primary keys of DIM_DIAGNOSTIC_TEST and DIM_DISEADE dimensions respectively. They include several attributes that would define diseases and several diagnostic tests respectively. Patient_Id, Date_of_Measurement_Id, Dis-

ease_Id, Time_of_Measurement_Id, and Test_Id deed as the composite primary key for Patient fact table. The DIM_IMAGE Dimension is attached to the FACT_PATIENT_IMAGE_DETAIL table, here Patient_Id and Image_Id with time_Id and date-Id deed as the composite primary key. The FACT_PATIENT_IMAGE_DETAIL integrate attributes that would record facts corresponding to the numerical conversion of medical images like area, perimeter, skewness, median gray value, mean gray value, diagnosis, etc.

4. SYSTEM DESCRIPTION

4.1. Medical Image Data Warehouse Architecture

The medical image data warehouse framework is based on a multitier architecture that divided into many phases, the data source systems, staging data system, data warehouse business service elements, and graphical doctor interface (GDI) framework into distinct layers. Figure4 illustrates the testbed performance of the breast cancer imaging data warehouse system.

4.1.1. Data source system. Source systems include various medical database and operational systems such as a HIS, PACS, RIS, breast-imaging database, breast cancer surgery database, pathology lab database, and psychology database from diverse research lab and medical departments of public hospitals in Tunisia. The data in our medical image data warehouse can be widely classified into breast-imaging modalities, numerical microscopic images, and results from pathology, diagnostic reports, Medical history Patient lifestyle, genetics data, Environmental, patient demographics, lab tests, surgical results, and post-treatment findings every year after surgery.

Patients undergo a series of examinations under well-defined protocols at the Tunisian National Institute of Public Health (INSP) and The oncology radiotherapy department of habib bourguiba university hospital (ORDHBUH) of sfax. Each breast image data set include a DICOM header, which itself includes textual data concerning the patient's information, name, age, date, imaging modality, scanning specification, attending specialist, and such. The generation and extraction of textual data from the DICOM header and the diagnostic textual reports can be automatic, though the extraction and segmentation of medical images often are determined interactively.
In case, breast cancer is diagnosed, the physician adds additional laboratory tests to help with prognosis. The two most ordinary tests are the HER2/neu test and the hormone receptor test [31] [32]. Outcomes from these tests can supply information into which cancer processing options may be most efficient for patients. The test outcomes are extracted from the networked HIS by an HL7 gateway. Software components retrieve the test result from the HIS reports and place them into the breast cancer data warehouse. The HL7 interface is founded on the commercial interface motor, The Rhapsody Messaging Toolkit.
The tumor surgery and cancerology databases reside on distinct Computers running Microsoft Access in the hospital-wide network. The data are first deposit into Microsoft Excel files and then automatically charged into the breast cancer data warehouse. The psychometric test results are manually recorded into the database by the Web from the paper archives, whilst pathology results and medical images are procured by FTP transfer.

4.1.2. Data Staging Area. Data staging area is the beginning depository and cleansing system for data that are transferring to the data warehouse server. They imply an ensemble of computational operations that clean, change, integrate, reduplicate, get ready, and release source data for design and utilize in the medical image data warehouse. It would be good if the data staging area were a unique centralized easing on one unit of hardware. In practice, after all, data arrive from diverse source systems on different platforms and in view of the requirements to use a diversity of software packages for diverse image processing trial, the data staging area of the data warehouse is divided over a set of computers (ACER Aspire E5-575G Intel Core i5-7200U Server NT).
The HL7 FHIR has been known as a very favorable draft standard for medical data exchange [33]. The popular Fast Healthcare Interoperability Resources (FHIR) is a fast-standard invented via Object Management Group (OMG) to help in distributed objects technology for exchanging data sources by component covering of HL7 interfaces. FHIR is going away in both method and product from previous HL7 standards such as HL7 V2 and HL7 3. The FHIR development process itself uses a progressive, iterative approach to improve the standard reflective of today's healthcare best practices for complex systems design. There is a deep concentrate by the FHIR development group on usability and suitability for purpose of the end-product [34].
The FHIR potential purpose to simplify and accelerate HL7 adoption by being facilely consumable but robust, and by employing open Internet standards where possible [34]. The image processing modules of the data staging area require and choose image processing algorithms for image filtering, segmentation of breast tissues, estimation, feature extraction from the segmented tumor region, quantitation, enhancement (for example image sharpening and contrast adjustment), classification, and structural or functional mapping, predication of tumor type into benign or malignant [35] [36]. An issue of data retrieval components provides access to lab test results and patient archives from the HIS and other province database systems. The data staging area also offers a temporary depository in both a relational database (Oracle 11.1) and a file management system disk space of 32 GB to clean, transform, join, and arrange data and images files.

For our present medical image data warehouse project, committed resources (research help) have been used to enhance routine and specify data loading and transformation tasks. For more refined data interpretation and image analysis, nevertheless, we often imply experienced end-

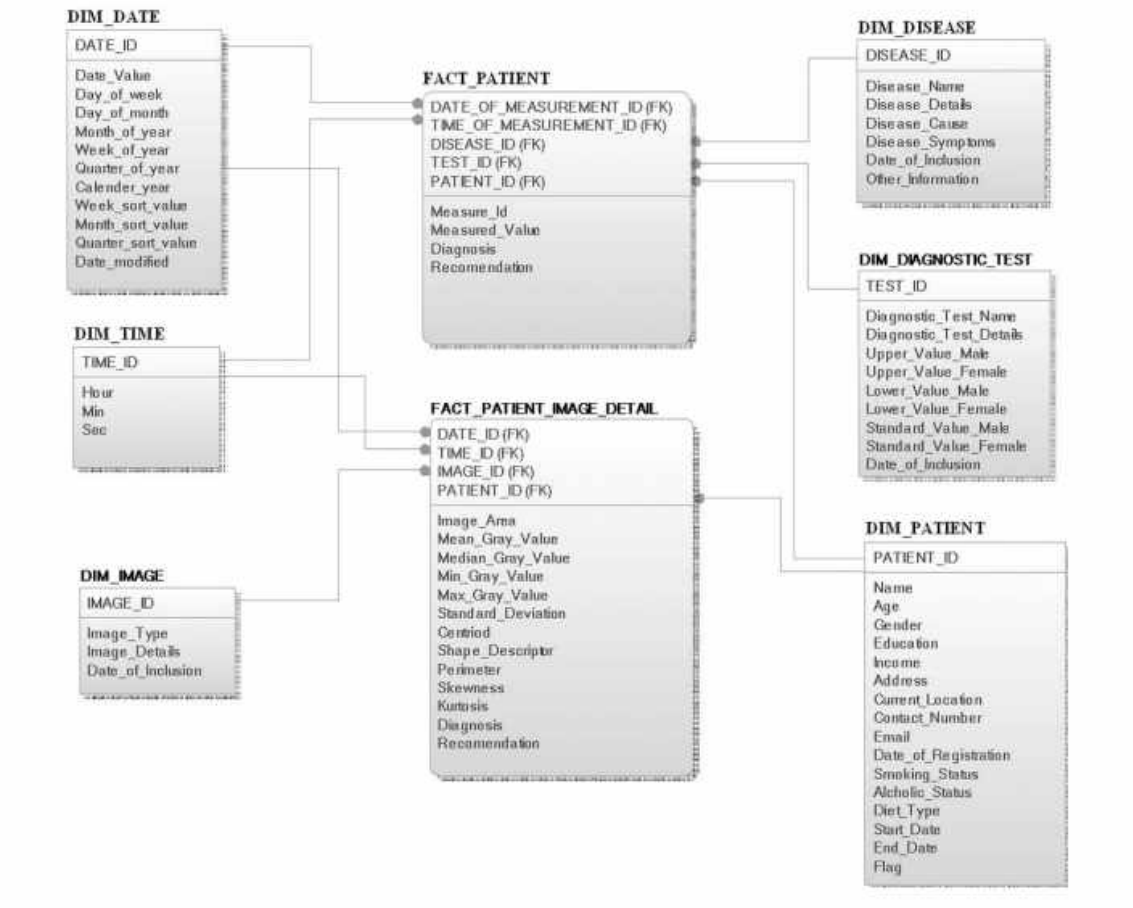


Figure 3. The star model of the breast-imaging data warehouse system.

users—e.g., oncologists, Responsible for women's imaging sector (sinology and anatomic-pathology), oncogenic, Psychologist, Nurse coordinator breast —during the staging process. They are willing to play an important part since they would obtain values out of the integrated breast data warehouse that they could not do so lonely.

4.2. Recording and Quantifying a Medical Image Data

An essential feature differentiating a medical image data warehousing system from traditional data warehousing system is the capacity of the first to manage, correlate, and query quantitative medical image data. Figure 6 illustrates the processing steps that appear in the data staging zone to extract quantitative breast image and textual features from the operational database' store data, from the RIS, PACS and HIS. The steps can be used for more imaging modalities.

Feature extraction is founded on the a prior approach, in difference to automatic and dynamic characteristic extraction all along the end-user query, as projected in several non-medical image database systems.For medical image data warehousing systems, co-registration of functional PET/scan is an important step toward combining functional information from PET scanner with anatomic information in MR scanner [37].

Various co-registration algorithms have been issued in the literature and are used in functional breast imaging studies. Image registration permits the combination of different types of functional information (such as PET/scan, scintigraphy imaging, More recently all modalities, fMRI, ultrasound) and structural information (such as MR images), setting the stage for feature extraction. Our information system interfaces with Automated Image Registration software application (Mirada Medical RTx) as well as much proprietary technique in commercial image analysis for registration of medical breast images, including PET/scan, x-ray mammography, US, and MRI. The correlated medical image data sets are ciphered into the intended data warehouse model to distribute as ultimate indexing in medical image database queries.

For example, recording the functional medical images of MRI with PET/scan of the actual patient permits the intrinsically good spatial resolution of MR images to be employed in quantitatively analyzing functional information (quantify the radiation functional adjusts of radiotherapy aims and

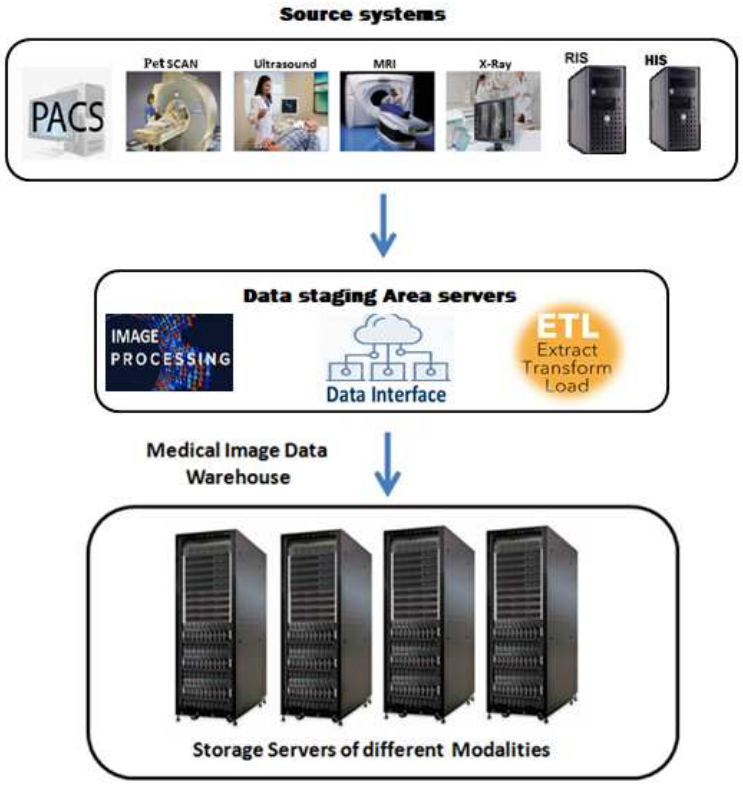


Figure 4. The multitier architecture of the breast-imaging data warehouse system.

censorious organs). The regions of interest derived from segmenting the mammogram's image are transformed into the IRM space for comparison [38].

The segmentation and extraction of medical images are done interactively using an environment for multidimensional image visualize and analysis VIDATM (volumetric image display and analysis) and Medical Breast Image software. Medical Image features are arranged into different levels of detail: Medical Image data set, anatomy, pathological structures, microscopic, and genetic system. We separated the medical features into primitive and logical. Primitive features immediately get from the medical images, such as shape, texture, volume of certain organs, and detection of invasive breast carcinoma and Ductal carcinoma in situ, in MRI or metabolic activities of breast tissue in PET/scans [39]

PET/scan calculates the variation in the metabolic rate. Just so, because cancer cells have a slightly important metabolic rate than other cells, the PET/scan will illustrate hot-spots of heightened density where malignant cancer tumors are probably developing. Since PET/scan method checks metabolic procedures at the molecular level, a fine-adjustment of the process employing dyes-engineered to respond only to special molecular features may be able to donate a clearer recognition of given breast cancer. A PET scan can also help in predicting cancer behavior and to assist in adapting the most efficient treatment [40].

Logical characteristics are a summary demonstration of medical images at diverse levels of detail and deeper field semantics that denote, for example, whether the volume of an anatomic unit is normal or whether some breast tissue is hypo metabolic in reference to some settled data. These logical characteristics are synthesized from primitive ones and supplemental domain knowledge.

The retrieve and extraction of medical data from the DICOM header and textual reports and can be an automatic process. Some keywords or phrases are automatically retrieved from the doctor's textual reports for indexing aims. All PACS Medical image files contain DICOM headers, which hold patient and imaging review information. The DICOM is formed by sequential data element tags, each including a group family number and component number. For example, the Patient Name value is situated in group 0001, component 0001. This value is automatically retrieved and transmitted into the medical data warehouse column PatientName.

TABLE 1. DATA COMPONENT TAGS ACCORDING TO THE DICOM HEADER

Family group	component	Name
0001	0001	PatientName
0001	0011	IDPatient
0022	0012	CivilState
0017	0014	RegionExamined
0006	0002	Pregnancies

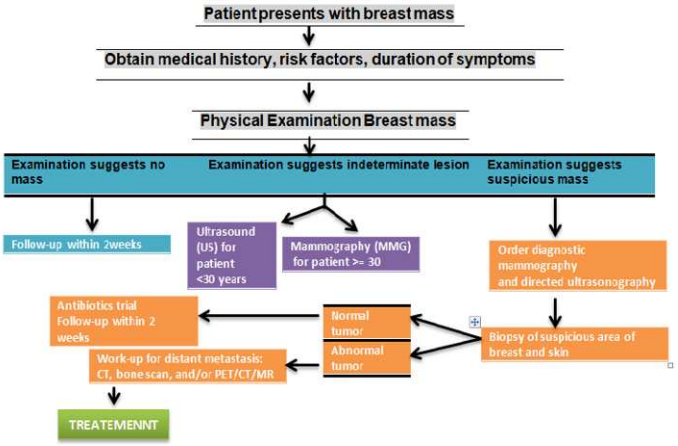


Figure 5. Diagnostic workup of Inflammatory Breast Cancer.

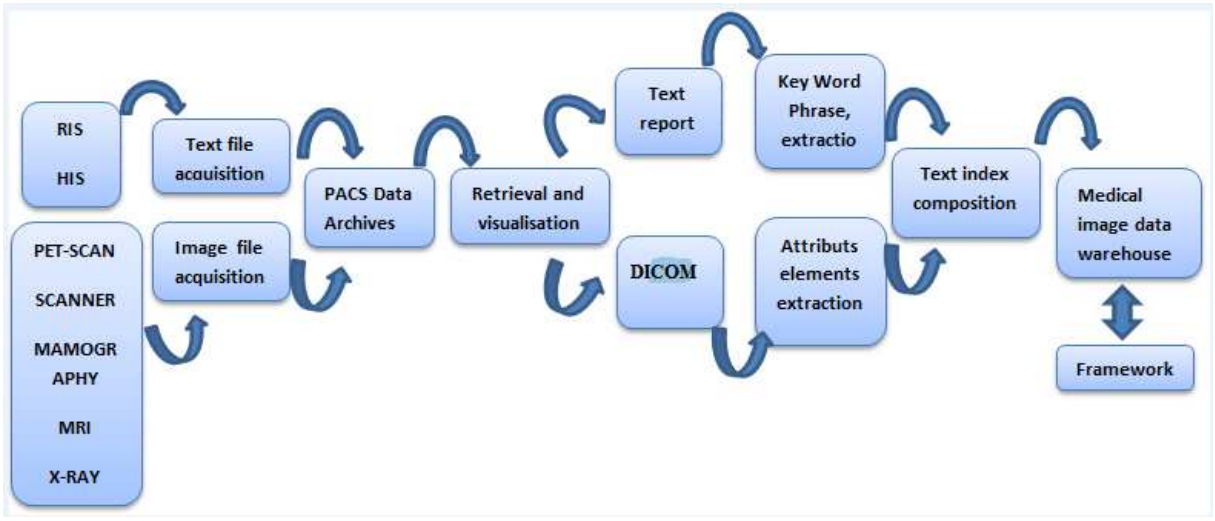


Figure 6. The operational flow of extracting the medical image and text characteristics into a medical image data warehouse for ulterior data analysis and mining. particular knowledge or heuristics is launched to help the query and navigation of the breast image data warehouse.

5. STATUS REPORT

We are utilizing the medical breast imaging data warehouse system to support several clinical treatments and research areas of cancer tumors. This part shortly explains two distinct mores of medical image data warehousing systems in encouraging clinical application and medical scientific research.

5.1. Diagnostic Workup of breast cancer Cases

Medical Image diagnostic workup is frequently a time consuming and hard-working task that implies gathering all pertinent medical records and images from several information systems and paper filing systems. Collecting and making a diagnostic workup, evaluation can from time to time endure weeks. The accessibility of an integrated medical data warehouse system that receives patient records and medical images for specific diseases supply a unique solution for faster and more veritable diagnostic workup. When a doctor uses a computer to launch an analysis session, step number one is to extract medical images and textual data from the PACS and diverse data sources. Figure 7 illustrates a screen that arises during the extraction of medical images and textual information from the medical data warehouse to the computer workstation. The Graphical User Interface (GUI) was created in the multimedia development environment (SQL developer from oracle, Sybase databases by downloading the jTDS driver) using orchestrates scripting language.

In the upper left of the user interface, the patient name is put (patient name and patient ID have been modified to preserve privacy). The system returns an explanation of the medical imaging studies matching the particularized patient name, and these items are visualized as rows in the upper scroll widget. Tapping on a row triggers

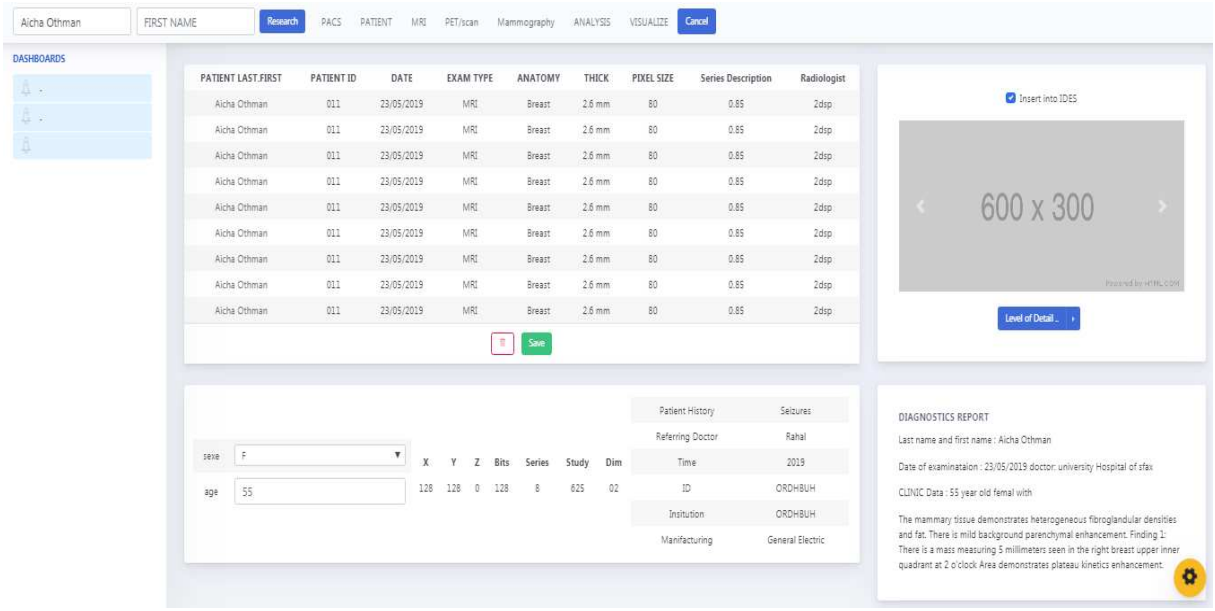


Figure 7. Multimedia user interfaces for information extraction in the breast image data warehouse.

extraction of the medical diagnostic report, and tapping on the "extract Image" button triggers the extraction of the selected medical image data set. Software units automatically transform the DICOM-founded image file to formats legible by the displaying packages on the computer workstation. With the medical imaging studies now existing in the workstation, medical information is quantitated, retrieved, and united in different ways to help in the localization task. First, within breast density assessment utilizing MRI, it is required to segment the breast orderly to donate total breast volume and eliminated non-breast surrounding tissues. One of the ultimate ideas for MRI automatic lesion capture is that surrounding non-breast regions such as the thoracic cavity, chest wall muscles, lungs, and heart must be before segmented out to stop false-positive detections on these regions, and a processing matrix is produced to co-register the volumetric PET/scan with the recently prepared MRI data set. The clinician can set diverse registration parameters, such as thresholds, sampling, initialization files, and iterations.

Exploiting the very good anatomic details of MRI, the region of interest viability of the breast-imaging data warehouse system is utilized to interactively segment breast tumor region. The resulting segmentation parameters are conjoint with the transformation matrix and FDG-PET to raise the positive predictive value (PPV) and particularity for patients in whom the MR results alone would be non-specific. Quantitative medical data, glucose estimate, and volumes cannot be acquired by traditional workups and are essential in determining the relevance of patients for surgical treatment.

After the treatment, the doctor records all the analyzed outcomes into the breast-imaging data warehouse for future studies. The potency of organizing the analyzed medical data in the breast-imaging data warehouse is the capability to-do online analytic processing of the collected medical data. The database can be interrogated on all number of keys, ranging from patient name to medical image features and textual report keywords. Complex queries—such as "Find female patients aged over 46 years with right breast area ¿ 120 cm3 corresponding density area ¿ 70 cm3 and density ¿ 50%"'can also be executed.

Names of matching patients are returned to the computer, and whole medical image data sets can be extracted by tapping on the thumbnail images. It is primordial to mark the capability to query the quantitative values of medical image features using the breast imaging data warehouse.

5.2. Medical Research Instance

Data Analysis and Decision Threshold on inflammatory breast cancer lateralization

Mammography is a used method of localizing the breast cancer zone. Mammography with IBC (inflammatory breast cancer) may demonstraionte a mass, calcification, or architectural distortion. The good contrast resolution of digital mammography permits visualization of trabecular and stromal thickening, and skin thickening, and broadcast raised breast density—returns that are often related to IBC. A principal mass lesion or a group of doubtful calcification is less communal in IBC than in non-IBC [**?**].

Thus, it is approved that women with suspected IBC suffer bilateral mammography, which will furnish the detection of the contralateral breast. Nevertheless, breast abnormalities occur periodically and may not be facilely captured. Magnetic resonance imaging may be approved in patients with doubtful IBC when mammography reveals

no breast parenchymal lesion. Through its employ is controversial, PET/CT is commonly employed for patients with IBC because early detection of far metastasis may ease the control of the metastatic disease. MR imaging examinations, and consistently PET scans, are used to provide additional information about breast cancer [41].
The oncology radiotherapy department of habib bourguiba university hospital(ORDHBUH) of sfax actually performs additional non-invasive breast-imaging examinations—i.e., PET —associated with clinical research, on a selected set of breast cancer patients. The multiple-medical image data sets from these patients are treated, organized, and designed for on-line medical data performance and analysis.

The breast-imaging data warehouse framework has been used to analyze breast images from 287 patients. In cases in which no breast abnormalities were detected on MRI, co-registered MRI was utilized to detect breast abnormalities on PET. Not founded abnormalities and wrong interpretations of independently elucidated high-resolution F18-fluorodeoxyglucose (18FDG)-PET images were rightly identified with co-registration to MR imaging, permitting detection of breast cancer tumor in patients [42]. The data analysis module of the image data warehouse is also used to investigate the lateralization concordance between various breast-imaging modalities in IBC. Such types of queries are priceless to define the clinical efficiency of the diagnostic procedure. Figure 8 and9 illustrate sample medical data analysis pages. In this application, surgery result is used as the gold standard for reference, and MRI and Pet/scan are compared with each other and with surgery results in patients who suffer surgical treatment more than a year previously .
The breast-imaging data warehouse permits users to personalize the presentation format (customize the display layout) in advance. In Figure 8 and 9, the user has picked pie charts and bar graphs to shows the analyzed results, rather than text formats or tables. Moreover, as demonstrated in Figure 8 and 9, the bottom, a detailed non-concordance listing in the lower-left medical text field permits the user to choose a particular patient case and then recuperate the rest of the breast cancer tumor record for detailed study.

The current approach to lateralization of IBC is qualitative and frequently depends on the experience of human imaging specialists. Decision thresholds are numeral thresholds for multimodality inflammatory breast cancer lateralization and placement of abnormalities zone, which leads to effective specific surgery. These thresholds are given by analyzing the breast-images of many patients for volume loss in MR imaging, and metabolism in PET.
We are actually developing an XML-based GUI to connect to familiar statistical packages, such as SAS and SPSS, for more effective data analysis and discovery. The analytic ability of such a breast-imaging data warehouse increase with the increase in treated patient images and medical instances.

6. Conclusion

The health domain is developing rapidly in the world. to permit the data centralization and take data advantage, we build a medical data warehouse for breast cancer, the data warehouse permits to integrate textual and image data. the star scheme consists to understand the tables and the relationships. We adapt the above data warehouse architecture in the context of our project; it is dedicated to the health care sector. Specifically for patient treatment at a Public Hospital. The data warehouse is a powerful and complete solution combining the ETL tool for integration, transformation, and loading of information in our data warehouse for public hospitals, and better mechanisms or decision-making applications currently available. The data warehouse responds to the needs of physicians, oncologists, pathologists, and other Health professional managers. It helps decision-makers in decision making. To facilitate access to information, and have the necessary and relevant data, we will use neural networks as an optimization technique. We will probe the breast image data warehouse to testing the generality of the medical framework by employing it to analyze other breast-imaging related disease progress and treatment methods, such as these for Metastatic Breast Cancer.

Acknowledgments

The authors would like to thank...

References

[1] Q. Li and R. Nishikawa, *Computer-Aided Detection and Diagnosis in Medical Imaging*, ser. Imaging in Medical Diagnosis and Therapy. CRC Press, 2015.

[2] R. Babić, Z. Milošević, i. Boris, and G. Stanković-Babić, "Radiology information system," *Acta medica medianae*, vol. 51, pp. 39–46, 12 2012.

[3] H. Huang, "Pacs-based multimedia imaging informatics: Basic principles and applications," in *TEMPLATE'06, 1st International Conference on Template Production*. Wiley, 2019.

[4] C. H. S. E. L. D. R. G. G. L. A. G. S. H. K. R. L. D. R. S. Foran David J, Chen, "Roadmap to a Comprehensive Clinical Data Warehouse for Precision Medicine Applications in Oncology," *Cancer Informaticsl*, vol. 16, pp. 1–10, 2017.

[5] J. R. Wright, "Pac contributions, lobbying, and representation," *The Journal of Politics*, vol. 51, no. 3, pp. 713–729, 1989.

[6] D. Haak, C.-E. Page, S. Reinartz, T. Krüger, and T. M. Deserno, "Dicom for clinical research: Pacs-integrated electronic data capture in multi-center trials," *Journal of digital imaging*, vol. 28, no. 5, pp. 558–566, 2015.

[7] C. R. Berger, M. Guerriero, S. Zhou, and P. Willett, "Pac vs. mac for decentralized detection using noncoherent modulation," *IEEE Transactions on Signal Processing*, vol. 57, no. 9, pp. 3562–3575, 2009.

[8] N. Stolba, M. Banek, and A. M. Tjoa, "The security issue of federated data warehouses in the area of evidence-based medicine," in *First International Conference on Availability, Reliability and Security (ARES'06)*. IEEE, 2006, pp. 11–pp.

[9] S. Singh, "Data warehouse and its methods," *Journal of Global Research in Computer Science*, vol. 2, no. 5, pp. 113–115, 2011.

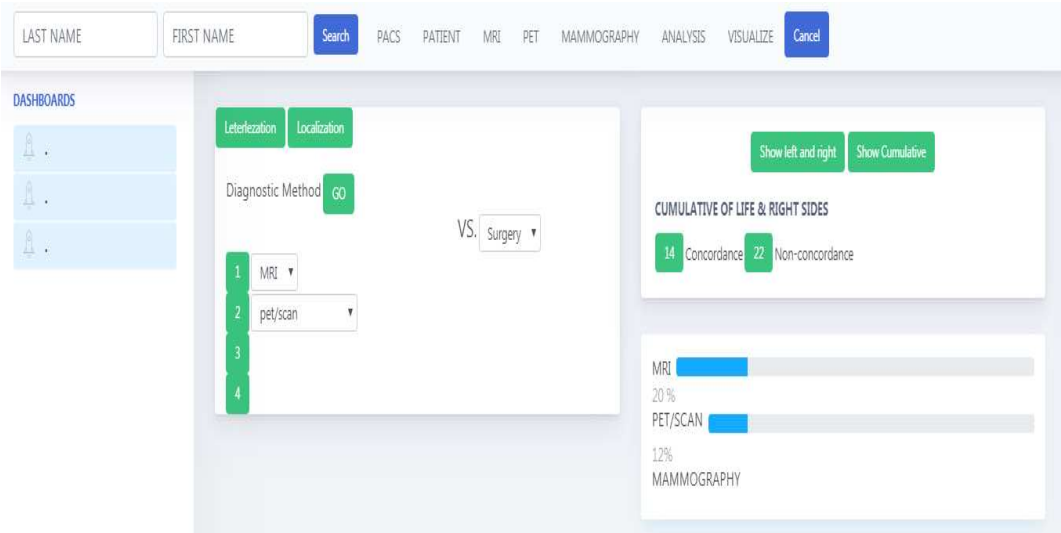


Figure 8. Data analysis pages of the breast-imaging data warehouse, illustrating the data analysis functions on lateralization and localization concordance.

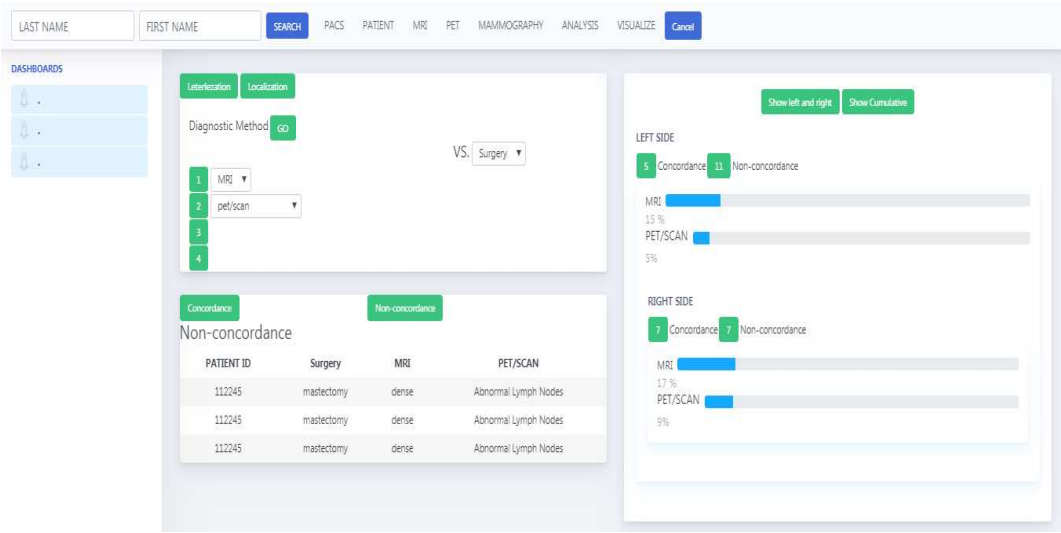


Figure 9. Data analysis pages of the breast-imaging data warehouse, illustrating the data analysis functions on lateralization and localization concordance.

[10] S. Timón, M. Rincón, and R. Martínez-Tomás, "Extending xnat platform with an incremental semantic framework," *Frontiers in Neuroinformatics*, vol. 11, p. 57, 2017.

[11] L. Ohno-Machado, V. Bafna, A. A. Boxwala, B. E. Chapman, W. W. Chapman, K. Chaudhuri, M. E. Day, C. Farcas, N. D. Heintzman, X. Jiang *et al.*, "idash: integrating data for analysis, anonymization, and sharing," *Journal of the American Medical Informatics Association*, vol. 19, no. 2, pp. 196–201, 2011.

[12] T.-T. Kuo and L. Ohno-Machado, "Modelchain: Decentralized privacy-preserving healthcare predictive modeling framework on private blockchain networks," *arXiv preprint arXiv:1802.01746*, 2018.

[13] P. McConnell, R. C. Dash, R. Chilukuri, R. Pietrobon, K. Johnson, R. Annechiarico, and A. J. Cuticchia, "The cancer translational research informatics platform," *BMC medical informatics and decision making*, vol. 8, no. 1, p. 60, 2008.

[14] K. Shimokawa, K. Mogushi, S. Shoji, A. Hiraishi, K. Ido, H. Mizushima, and H. Tanaka, "icod: an integrated clinical omics database based on the systems-pathology view of disease," in *BMC genomics*, vol. 11, no. 4. BioMed Central, 2010, p. S19.

[15] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson *et al.*, "The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," 2012.

[16] A. Tan, B. Tripp, and D. Daley, "Brisk—research-oriented storage kit for biology-related data," *Bioinformatics*, vol. 27, no. 17, pp. 2422–2425, 2011.

[17] F. Scheufele, B. Wolf, M. Kruse, T. Hartmann, J. Lempart, S. Mühlich, A. F. Pfeiffer, L. J. Field, M. J. Charron, Z.-Q. Pan *et al.*, "Evidence for a regulatory role of cullin-ring e3 ubiquitin ligase 7 in insulin signaling," *Cellular signalling*, vol. 26, no. 2, pp. 233–239, 2014.

[18] Y. Schwartz, A. Barbot, B. Thyreau, V. Frouin, G. Varoquaux, A. Siram, D. Marcus, and J.-B. Poline, "Pyxnat: Xnat in python," *Frontiers in neuroinformatics*, vol. 6, p. 12, 2012.

[19] D. Solodovnikova, L. Niedrite, and N. Kozmina, "Handling evolving data warehouse requirements," in *East European Conference on Advances in Databases and Information Systems*. Springer, 2015, pp. 334–345.

[20] D. Solodovnikova, "Data warehouse evolution framework."

[21] S. Z. Oliva and J. C. Felipe, "Optimizing public healthcare management through a data warehousing analytical framework," *IFAC-PapersOnLine*, vol. 51, no. 27, pp. 407–412, 2018.

[22] M. A. Levin, T. T. Joseph, J. M. Jeff, R. Nadukuru, S. B. Ellis, E. P. Bottinger, and E. E. Kenny, "igas: A framework for using electronic intraoperative medical records for genomic discovery," *Journal of biomedical informatics*, vol. 67, pp. 80–89, 2017.

[23] "Conception et applications d'un framework d'entrepot de donnees d'images a multimodalites," *Journal de l'American Medical Informatics Association*, vol. 9, pp. 239–254.

[24] B. Parmanto, M. Scotch, and S. Ahmad, "A framework for designing a healthcare outcome data warehouse," *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, vol. 2, 2005.

[25] M. Chen, J. Yang, J. Zhou, Y. Hao, J. Zhang, and C. Youn, "5g-smart diabetes: Toward personalized diabetes diagnosis with healthcare big data clouds," *IEEE Communications Magazine*, vol. 56, no. 4, pp. 16–23, April 2018.

[26] M. Smith, L. L. Roos, C. Burchill, K. Turner, D. G. Towns, S. P. Hong, J. S. Jarmasz, P. J. Martens, N. P. Roos, T. Ostapyk, J. Ginter, G. Finlayson, L. M. Lix, M. Brownell, M. Azimaee, R.-A. Soodeen, and J. P. Nicol, *Health Services Data: Managing the Data Warehouse: 25 Years of Experience at the Manitoba Centre for Health Policy*. New York, NY: Springer US, 2019, pp. 19–45.

[27] H. Aldosari, B. Saddik, and K. A. Kadi, "Impact of picture archiving and communication system (pacs) on radiology staff," *Informatics in Medicine Unlocked*, vol. 10, pp. 1 – 16, 2018.

[28] S. e. N. S. Jukic, Nenad et Vrbsky, *Systèmes de bases de données: Introduction aux bases de données et aux entrepôts de données*. Prospect Press, 2016.

[29] T. T. T. A. D. B. N. Kunjan, K., "A multidimensional data warehouse for community health centers," *AMIA ... Annual Symposium proceedings*, 2015.

[30] R. Kimball and M. Ross, *The Kimball group reader: relentlessly practical tools for data warehousing and business intelligence*. John Wiley Sons, 2010.

[31] C. Yeung, J. Hilton, M. Clemons, S. Mazzarello, B. Hutton, F. Haggar, C. L. Addison, I. Kuchuk, X. Zhu, K. Gelmon, and A. Arnaout, "Estrogen, progesterone, and her2/neu receptor discordance between primary and metastatic breast tumours—a review," *Cancer and Metastasis Reviews*, vol. 35, no. 3, pp. 427–437, Sep 2016.

[32] S. Rossi, M. Basso, A. Strippoli, V. Dadduzio, E. Cerchiaro, R. Barile, E. D'Argento, A. Cassano, G. Schinzari, and C. Barone, "Hormone receptor status and her2 expression in primary breast cancer compared with synchronous axillary metastases or recurrent metastatic disease," *Clinical Breast Cancer*, vol. 15, no. 5, pp. 307 – 312, 2015.

[33] J. C. Mandel, D. A. Kreda, K. D. Mandl, I. S. Kohane, and R. B. Ramoni, "Smart on fhir: a standards-based, interoperable apps platform for electronic health records," *Journal of the American Medical Informatics Association*, vol. 23, no. 5, pp. 899–908, 2016.

[34] D. Bender and K. Sartipi, "Hl7 fhir: An agile and restful approach to healthcare information exchange," in *Proceedings of the 26th IEEE international symposium on computer-based medical systems*. IEEE, 2013, pp. 326–331.

[35] J. C. Russ, *The image processing handbook*. CRC press, 2016.

[36] A. Sebaa, F. Chikh, A. Nouicer, and A. Tari, "Medical big data warehouse: Architecture and system design, a case study: Improving healthcare resources distribution," *Journal of Medical Systems*, vol. 42, no. 4, p. 59, Feb 2018.

[37] A. Kenneth, *Skeletal Circulation In Clinical Practice*. World Scientific Publishing Company, 2016.

[38] S. C. e. F. M. K. e. K. A. H. e. S. H. C. e. S. B. e. v. S. G. K. e. K.-H. R. A. Goerres, Gerhard W et Michel, "Suivi des femmes atteintes d'un cancer du sein: comparaison entre l'irm et la tep au fdg."

[39] O. Humbert, "Imagerie tep au 18f-fdg du cancer du sein: étude du comportement métabolique des différents phénotypes tumoraux et prédiction de la réponse tumorale à la chimiothérapie néoadjuvante," Ph.D. dissertation, 2015.

[40] R. K. e. S. J. M. e. G. J. R. e. E. G. K. e. L. R. B. e. L. H. M. e. G.-V. K. e. K. B. F. e. S. E. K. e. a. Dunnwald, Lisa K et Doot, "Metabolisme de la tumeur pet chez les patientes atteintes d'un cancer du sein localement avance et recevant une chimiotherapie neoadjuvante interet des mesures statiques et cinetiques de l'absorption du fluorodesoxyglucose," *Recherche clinique sur le cancer*, vol. 17, no. 8, pp. 2400–2409, 2011.

[41] J. C. Gooch and F. Schnabel, *Inflammatory Breast Cancer*. Cham: Springer International Publishing, 2019, pp. 105–108.

[42] A. N. Melsaether, R. A. Raad, A. C. Pujara, F. D. Ponzo, K. M. Pysarenko, K. Jhaveri, J. S. Babb, E. E. Sigmund, S. G. Kim, and L. A. Moy, "Comparison of whole-body 18f fdg pet/mr imaging and whole-body 18f fdg pet/ct in terms of lesion detection and radiation dose in patients with breast cancer," *Radiology*, vol. 281, no. 1, pp. 193–202, 2016, pMID: 27023002.

Recommender systems based on detection in academic social network

Smail Boussaadi, Dr Hasina Aliane , Prof Ouahabi Abdeldjalil

Dihia Houari , Malika Djoumagh.

Abstract— The speed with which new scientific articles are published and shared on academic social networks generated a situation of cognitive overload and the targeted access to the relevant information represents a major challenge for researchers. In this context, we propose a scientific article recommendation approach based on the discovery of thematic community structures, it focuses on the topological structure of the network combined with the analysis of the content of the social object (scientific article), a strategy that aims to mitigate the cold start problems and sparcity data in scoring matrix. A key element of our approach is the modeling of the researcher's thematic centers of interest derived from his corpus (a set of articles that interested him). In this perspective we use the technique of semantic exploration and extraction of latent topics in document corpora, LDA(Latent DirichletAllocation), an unsupervised learning method which offers the best solution of scalability problemcompared to other techniques of topic modeling. this technique allows us to build a profile model in the form of vectors in which the components are the probabilistic distributions on topics that reflect the interests of the researcher. The profile models thus constructed will be grouped into thematic clusters based on dominant topics using the fuzzy clustering algorithm, since the same topic can be treated in different scientific fields. Will follow a step of detection of community structures in thematic clusters to identify significant communities, the aim of this step is to project the recommendation process in a small space allowing better performance by reducing the computation time and the storage space for researcher / article data. The preliminary results of the experience of our approach on a population of 13 researchers and 60 articles shows that the articles generated by the recommendation process are very relevant to the target researcher or his community.

Index Terms— LDA , Recommender system, academic social network, community detection, the fuzzy clustering algorithm.

1 INTRODUCTION

Academic social networks like generalist social networks provide millions of researchers with functionalities that allow them to promote their publications, to find relevant articles and to discover trends in their areas of interest. However, the rapidity with which new articles are published and shared, especially on these academic social networks, generates a situation of cognitive overload and is therefore a major challenge for the researcher in search of relevant and recently published information. It is in this context that scientific article recommendation systems are used to filter the huge amount of articles shared on these platforms. In recent years many researchers have become interested in recommending scientific articles. [1][2][3][4][5].

The recommendation of scientific articles, aims at recommending relevant articles in correlation with the interests of a researcher or a group of researchers. However, due to their ever-increasing size, the analysis of academic social networks, on which a scientific article recommendation system is based, has become a complex task, and one of the strategies adopted in the scientific literature to overcome this difficulty is the partitioning of the initial network into smaller subgroups, for example, researchers sharing the same areas of interest will be grouped into thematic communities, a strategy that reduces processing time and storage space since the observations relate to a group of researchers interested in the same themes, hence the interest of the community concept, which provides a relevant analytical framework for Understanding the collective and organizational dynamics of the overall network [6].

The essential of the literature on recommending systems shows that there is a consensus on the Classification of Recommending Systems for the three categories : methods based on collaborative filtering, content-based methods and hybrid methods [8][9][10][11].

Our article is organized as follows

In section 2 we present the bulk of the literature on recommendation approaches based on theme modeling and thematic community schemes, in section 3 we present our recommendation approach that aims to mitigate the problem of cold start and data sparcity. In order to evaluate the effect of taking into account theme modeling in the identification of thematic communities in academic social networks we conducted experiments on a real data set that we have build around 13 researchers and 60 articles, the remaining part of our approach, i.e. recommendation generation, will be the subject of a forthcoming article.

1.1 Content-Based Methods (CBF) : The content-based filtering process essentially takes place in two phases. The learning of the profile vector of the active researcher based on the history of his activity, There are different methods for this, including LDA (Latent Dirichlet Allocation). [12] (section 2.3). And the representation as a feature vector of the candidate article, defined by models such as TF-IDF which produces a weighted vector of terms [13] or sentence extraction that produces a description of the content of each candidate article in the form of a list of key words that reflect the essence of the topics addressed in each candidate article [14]. The second phase uses a similarity function that takes as input the profile vector of the active searcher and the vector representing the candidate article and provides a prediction score. Generally the similarity function is the cosine of the angle formed by the two vectors in question, this last phase produces a ranked list of articles whose top_N articles will be recommended to the target researcher [15]. In the field of recommending scientific articles, content-based filtering is the most widely used [16].

1.2 Methods based on collaborative filtering (CF)
Collaborative filtering is based on the sharing of opinions and evaluations between researchers on certain articles. The underlying idea is: if a researcher A evaluates or rates a U paper in the same way as another researcher B, if both researchers A and B have previously enjoyed other articles in a similar way. And unlike content-based filtering, collaborative filtering is independent of the content of candidate articles [17], [18]. Typically, researchers evaluations of articles are represented by a scoring matrix, where each line corresponds to a researchers evaluation history. Recommendations produced in a collaborative filtering process are based on similarity between researchers.

According to [19][20] there are two classes for collaborative filtering, memory-based methods and model-based methods.

Memory-based methods : this method exploits the entire usage matrix (Fig) to generate recommendations. The term memory (neighborhood) refers to users as well as articles. Thus the algorithms of this method can be divided into two categories: user based or items based.
User based: introduced by GroupLens [21], The principle of this method is to first determine which users are similar to the active user, which is equivalent to estimating the similarities between the row in the active user usage matrix with all other rows [22][23], then fills in the empty cells of the usage matrix with a prediction score. The calculation of similarity between the vectors representing the users is measured by the cosine, or Pearson's coefficient, The latter is the most widely used and the most efficient in terms of predictive accuracy [23].

items based: The principle of this method consists in predicting the appreciation of the active user for a candidate article, based on the assessments of the active user for articles similar to the candidate article[24]. The determination of similar articles can be calculated by the cosine of the article attribute vectors or the Pearson's coefficient.
Model-based methods:These methods are based on learning machine techniques such as probabilistic models (naive Bayesian classifier) [25] , clustering, and the most popular, latent factor models [26]. In order to address the shortcomings of memory-based collaborative filtering [27]. A model-based process learns to recognize complex patterns on offline training data so that it can generate predictions on test data or real-world online data.
A probabilistic model consists of calculating the probability $P(a|b,i)$ that the user a assigns the score b to the item i knowing its previous scores. The prediction pred(b; i) matches, either to the rating with the highest probability, or to the expected rating, as defined by the formula [26].

$$pred(a,i) = \sum_{b \in B} b.P(b \mid a,i)$$

B is the set of values that a note can take.
The clustering technique is often used as an intermediate step to bring together researchers sharing the same areas of interest or to group articles addressing the same topics into clusters, which will be exploited for further processing.

1.3 Hybrid Methods
A hybrid recommendation process combines both content-based and collaborative filtering techniques to increase recommendation performance[27].

2 RELATED WORK

Our work is linked to two main lines of research

1. Modelling of researcher profiles based on the LDA scheme.
2. The discovery of thematic communities in an academic social network, for the recommendation of scientific articles.

We review the relevant literature dealing with these two areas.

.

2.1 Modelling of researcher profiles based on the LDA scheme

The subject modeling is a powerful and practical tool for semantic exploration and subject extraction. One of the methods that has been the subject of many scientific publications in the field of recommendation systems, in particular the recommendation of scientific articles, is the LDA generative probabilistic model. Table .1 (see annex) presents a selection of relevant publications.

2.2 The discovery of thematic communities in an academic social network, for the recommendation of scientific articles.

The detection of communities of interest consists in identifying the best possible graphical partitioning of a network. In the context of our study this translates into the identification of subsets of researchers grouped together on the basis of similarity of thematic interest without any explicit social interaction between them. Recommendation approaches based on community structures can significantly limit the number of

users in the process of calculating the prediction and the results are more relevant given that the size of the data is limited to community members only [28]. In the literature several taxonomies are proposed for community detection algorithms, including the exhaustive study carried out by [29].
The majority of community detection methods have focused on the topological structure of the graph without analyzing the content exchanged between users. However, some approaches use content analysis through subject modeling techniques such as LSA[30] pLSA[31] LDA[12] , These approaches of community detection do not take into account the explicit links between network members. In their research, [32] combined topic modeling with link structure. And [33] proposed a framework to apply a semantically structured approach to the web service community modeling and discovery. Table .2 (see annex) presents a relevant selection of publications on the detection of thematic communities in academic social networks.

2.3 LDA model for the representation of a researchers profile

A researchers profile reflects his thematic interests, it is generated from its corpus, by automatic modeling of themes, a widely used technique for semantic exploration and theme extraction in large volumes of textual documents. In this perspective we apply the LDA model, An unsupervised learning technique that treats an article as a vector of words to identify unobservable themes. In the last decade many articles have addressed the use of the LDA model in recommending scientific articles[34][35][36][37].

Considering an article d composed of N words :
$d_i = \{w_1, w_2, ...w_N\}$

There is then a probabilistic relationship between the words w_i, the topics noted z_k(k∈ℕ) and the article in question d :

P($z_k|d$) : the probability of topic z_k in the document d
P($w_i|$ z_k) :the probability of word w_i in topic z_k;

The hyperparameters of the model:
α : represents the density of subjects in a document.

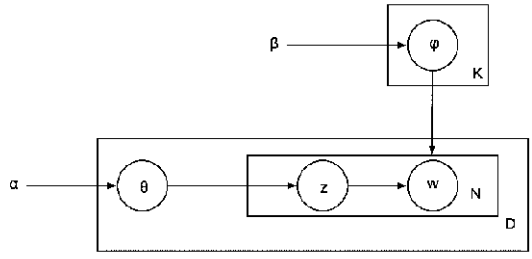


Fig. 1 graphic representation of the LDA model

B:represents the density of words in a subject.
K : number of topics
D : number of words in document d
N : number of documents.
The latent variables :

- θ document topic distribution.
- z word topic assignment.
- w observed word.

The modeling represented in the figure will produce the following results in the form of matrices :

- Topics x mots.
- Document x topics.
- Document x mots.

The model has two parameters to infer from the observed data, which are the distributions of the latent variables θ (document-subject) and z (subject-word). By determining these two distributions, it is possible to obtain the topics of interest on which researchers write.

3 Proposed Approach

In this section, we propose our approach to recommending scientific articles in order to solve the cold start problem that occurs in recommendation systems due to the lack of information on the one hand about a new user who has not yet interacted with the scientific articles, such as publishing an article, downloading, sharing,...and whose corpus is considered to be empty and on the other hand about a new article that has not been the subject of interest from researchers. Our approach, unlike the approaches described in section 3.1, is based on the modelling of subjects to build very precise profiles taking into account only the thematic area of interest of the researchers without taking into account other types of information that do not bring precision to the profile and may overload it.The problem of sparcity, particularly in the rating matrices, is dealt with by reducing the size of the space (researchers, articles) to a very small space (researchers, topics), and from the latter further reduces the size of the space to communities of thematic interest, which restricts the storage space and the processing time of the recommendation process to the neighbours of the active researcher only.

3.1 Description of the approach

we propose to carry out our approach with the following steps.Fig.4

Step 1: acquisition of data on the researcher by the web crawler technique.Fig.2 Indeed a crawling tool extracts all the information relating to the articles in the browsing history, the annotated tags that contain summaries of interesting articles, ... this step produces a schema (researcher, corpus), will follow a preprocessing for the construction of a dataset for the experimentation phase.

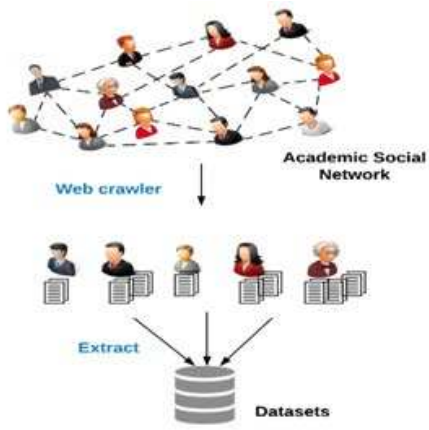


Fig. 2 construction of our datasets

Step 2: we apply the LDA model learning on the corpus of each researcher, to extract the different themes that constitute the thematic areas of interest for each researcher. Thus for each researcher we have his profile in the form of a probability vector on the subjects. We prepare for the next step, a matrix whose lines are the profile vectors of the researchers.

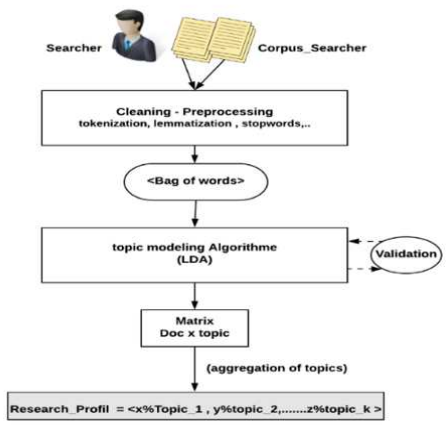


Fig. 3 Profile modeling

Step 3: Researchers who have a high probability for the same themes will form thematic clusters, by applying a flexible clustering algorithm.(fig.6)

Step 4: An algorithm is applied to detect thematic communities of interest, starting from the graph of res earchers weighted by a similarity between the profiles of the corresponding researchers. Thus, the neighbourhood of a target researcher will be identified among similar researchers in his community.(fig.7)

Step5: for a candidate article for recommendation on the basis of a prediction score, a correlation function will be applied between the profile of the target researcher and the vector representing the article in question in the form of a subject probability vector. for a set of candidate articles for recommendation, a ranking of the correlation scores in descending order

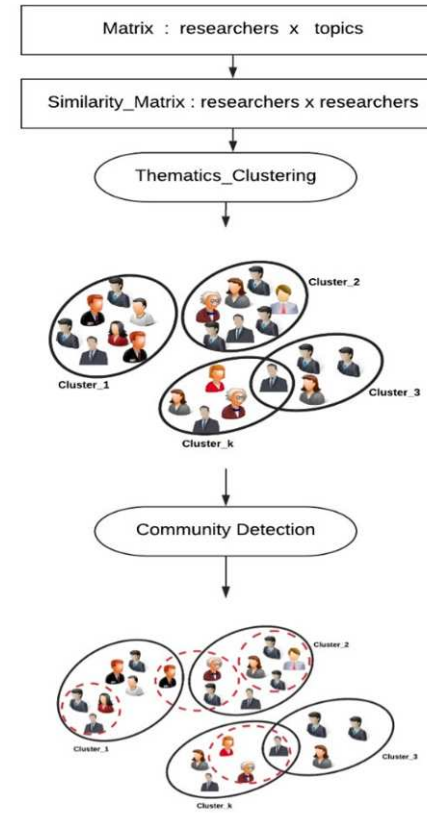


Fig.4 Presentation of the main stages of our approach

is produced and the articles with the highest scores will be consid-ered relevant for recommendation.

4 EXPERIENCE ET ANALYSE

we have represented our network of researchers by a graph weighted by the cosine similarity between the different entities.(fig.5) .
Each node represents a researcher and his identifier .

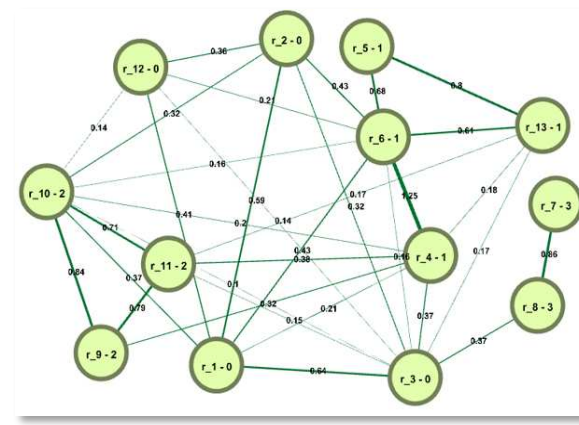


Fig. 5 Weighted graph researchers x researchers

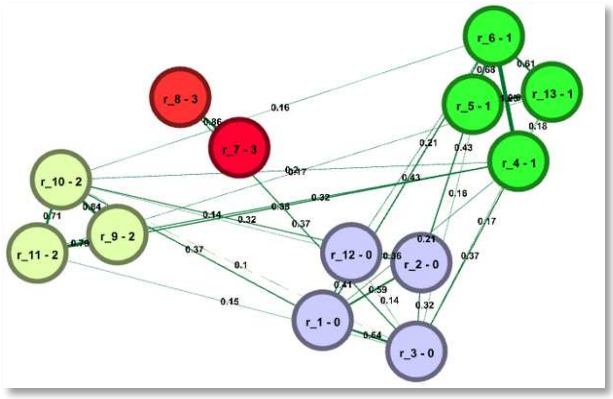


Fig. 6 Discovery of thematic communities (classes)

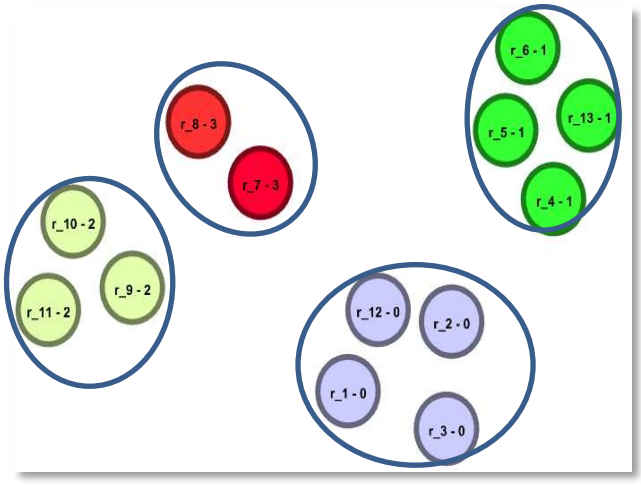


Fig. 7 ultimate thematic communities

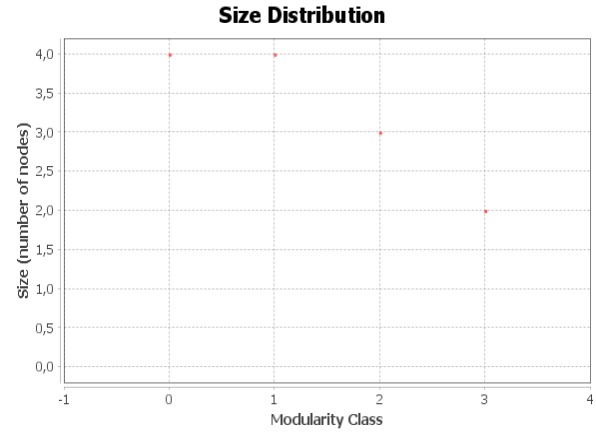


Fig.8 Modularity_score

The communities obtained after applying Blondels algorithm [40] with the Gephi tool produced the following classes : fig.8 Results :

- Modularity :0.366
- Modularity with resolution : 0.366
- Number of communities : 4

The communities discovered :

Community_0 ={researcher_1, researcher_2, researcher_3, researcher_12}.
Community_1 ={ researcher_4, researcher_5, researcher_6, researcher_13}.
Community_2 ={ researcher_9,researcher_10, researcher_11}.
Community_3 ={ researcher_7, researcher_8}.

5 Conclusion and Future Work

In this article we have proposed a scientific article recommendation approach based on the discovery of thematic communities of interest in the context of an academic social network.Taking into account only the subjects to model the profiles of the researchers made it possible to partition the network on the basis of similarity between researchers sharing a high probability for the same subject. Then a community detection algorithm is applied on all the clusters formed during the previous step to identify ultimate communities around thematic interests. The results we have obtained up to step 4 have shown that the inclusion of subjects in the modeling of researcher profiles with abstraction of other social information has provided very promising results so we plan to apply a process of recommending scientific papers for a targeted researcher and for these community neighbours.

ANNEX

TABLEAU 1 SELECTION OF RELEVANT ARTICLES ON LDA AND RECOMMENDATION OF ARTICLES

Authors	year	Articles
Jelodar et al	2019	Natural Language Processing via LDA Topic Model in Recommendation Systems.
Kim, S., Gil, J	2019	Research paper classification systems based on TF-IDF and LDA schemes
Amami et al	2017	A graph based approach to scientific paper recommendation.
Sahijwani, H. et al	2017	User Profile Based Research Paper Recommendation.
Dai et al	2017	Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network.
Younus et al	2014	Utilizing Microblog Data in a Topic Modelling Framework for Scientific Articles' Recommendation.
Kim et al	2013	TWITOBI: A Recommendation System for Twitter Using Probabilistic Modeling
Li et al	2013	Scientific articles recommendation.
Wang et al	2011	Collaborative Topic Modeling for Recommending Scientific Articles
Chenguang et al	2010	Research Paper Recommendation with Topic Analysis

TABLEAU 2 SELECTION OF RELEVANT ARTICLES ON THE DETECTION OF COMMUNITIES AND IN SOCIAL NETWORKS

Author	year	Articles
Alfarhood et al	2019	Collaborative Attentive Autoencoder for Scientific Article Recommendation
Xia et al	2019	Scientific Article Recommendation: Exploiting Common Author Relations and Historical Preferences
Sun et al	2018	A hybrid approach for article recommendation in research social networks
Wan et al	2018	HAR-SI: A novel hybrid article recommendation approach integrating with social information in scientific social network
Mao et al	2018	Academic Social Network Scholars Recommendation Model Based on Community Division
Wang	2016	Academic Paper Recommendation Based on Community Detection in Citation-Collaboration Networks
Reihanian et al	2016	Topic-oriented Community Detection of Rating-based Social Networks
Sobharani et al	2013	Application of clustering to analyze academic social networks
Zhao et al	2012	Topic oriented community detection through social objects and link analysis in social networks
Ding et al	2011	Community detection: topological vs. topical

REFERENCES

[1] Amami, M., Pasi, G., Stella, F., & Faiz, R. (2016). An LDA-based approach to scientific paper recommendation. Paper presented at the International Conference on Applications of Natural Language to Information Systems.

[2] Sugiyama, K., & Kan, M.-Y. (2015). A comprehensive evaluation of scholarly paper recommendation using potential citation papers. International Journal on Digital Libraries, 16(2), 91-109

[3] Beel, J., Gipp, B., Langer, S., & Breitinger, C. (2016). Research-paper recommender systems: A literature survey. International Journal on Digital Libraries, 17(4), 305-338.

[4] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong and F. Xia, "Scientific Paper Recommendation: A Survey," in IEEE Access, vol. 7, pp. 9324-9339, 2019.

[5] Cohendet, P., Creplet, F., Dupouët, 0 .. 2000.«Communities of practice and epistemic communities: a renewed approach of organizational learning within the firm ». Proceedings of the 5th Workshop on Economies with Heterogeneous Interacting Agents (WEHIA). Marseille. 15-17 June.

[6] Luc Brunet et André Savoie « la face cachee de l'organisation » edition Presses de l'Université de Montréal 2003.

[8] Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17(6), 734-749.

[9] Burke, R. (2000). Knowledge-Based Recommender Systems. Dans Kent, A. (Éd.),Encyclopedia of Library and Information Systems (Vol. 69): Marcel Dekker

[10] Resnick, P., & Varian, H. R. (1997). Recommender systems. Commun. ACM, 40(3), 56-58.ReutersThomson. (2008).http://www.thomsonreuters.com/content/PDF/scientific/Web_of_Knowledge_factsheet.pd

[11] Schafer, J. B., Konstan, J., & Riedl, J. (1999). Recommender systems in e-commerce,Proceedings of the 1st ACM conference on Electronic commerce. Denver,
Colorado, United States: ACM.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res.,3 :993–1022, March 2003b. ISSN 1532-4435

[13] P. Jomsri, S. Sanguansintukul, W. Choochaiwattana, "A framework for tag-based research paper recommender system: An ir approach", Proc. IEEE 24th Int. Conf. Adv. Inf. Netw. Appl. Workshops (WAINA), pp. 103-108, 2010.

[14] C. Caragea, F. A. Bulgarov, A. Godea, S. D. Gollapalli, "Citation-enhanced keyphrase extraction from research papers: A supervised approach", Proc. Conf. Empirical Methods Natural Lang. Process., pp. 1435-1446, 2014.

[15] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions," in IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, June 2005.

[16] Joeran Beel, Bela Gipp, Stefan Langer & Corinna Breitinger Research-paper recommender systems: a literature survey Erschienen in: International Journal on Digital Libraries (2016), 4. - S. 305-338 - doi.org/10.1007/s00799-015-0156-0

[17] Breese J.-S., Heckerman D., Kadie C., Empirical Analysis of Predictive Algorithms for Collaborative Filtering, Proceedings of the 14th Conference on Uncertainty In Artificial Intelligence (UAI'98), Wisconsin, USA, 1998, p. 43-52

[18] Goldberg D., Oki B., Nichols D., Terry D.-B., Using Collaborative Filtering to Weave an Information Tapestry, Communications of the ACM, vol. 35 (12), 1992, p. 61-70.

[19] Bell, R., & Koren, Y. (2007). Scalable collaborative filtering with jointly derived neighborhood interpolation weights. 2007 Seventh IEEE Int. Conf. on Data Mining (pp. 43-52). Washington, DC, USA: IEEE Computer Society.

[20] Aciar, S., Zhang, D., Simoff, S., & Debenham, J. (2007). Informed Recommender: Basing Recommendations on Consumer Product Reviews. IEEE Intelligent Systems 22, 39-47.

[21] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: an Open Architecture for Collaborative Filtering of N
etnews. In Proceedings of the ACM conference on computer supported cooperative work,1994

[22] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems : A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering, 17 :734–749, 2005.

[23] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, The Adaptive Web, number 4321 in Lecture Notes in Computer Science, pages 291–324. Springer Berlin Heidelberg, 2007.

[24] J. Beel, B. Gipp, S. Langer, L. Breitinger, "Research-paper recommender systems: A literature survey", Int. J. Digit. Libraries, vol. 17, no. 4, pp. 305-338, 2016.

[25] Jun Wang, Arjen P De Vries, and Marcel JT Reinders. 2008. Unified relevance models for rating prediction in collaborative filtering. ACM Transactions on Information Systems (TOIS) 26, 3 (2008), 16.

[26] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 448–456.

[27] T. Hofmann, "Latent semantic models for collaborative filtering," ACM Transactions on Information Systems, vol. 22, no. 1, pp. 89–115, 2004.

[28] J. Han and M. Kamber. Data Mining: Concepts and Techniques, 2001

[29] Fortunato S. Community detection in graphs. Physics Reports.2010;486(3):75.doi:10.1016/j.physrep.2009.11.00

[30] Deerwester, Caroline du Sud, Dumais, ST, Landauer, TK, Furnas, GW et Harshman, RA 1990. Indexation paranalyse sémantique latente. JASIS, 41 , 391-407.

[31] SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrievalAugust 1999 Pages 50–57https://doi.org/10.1145/312624.312649

[32] Zhu, Y., Yan, X., Getoor, L., & Moore, C. 2013. Scalable textand link analysis with mixed-topic link models. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD). Publishing, pp. 473–481.

[33] Zhao, A., & Ma, Y., 2012. A Semantically Structured Approach to Service Community Discovery,Semantics, Knowledge and Grids (SKG), 2012 Eighth International Conference on. Publishing, pp. 136-142.

[34] Schafer, J. B., Frankowski, D., Herlocker, J., and Sen, S.
(2007). Collaborative filtering recommender systems. In The adaptive web, pages 291–324. Springer

[35] A. Tsolakidis, E. Triperina, C. Sgouropoulou, N. Christidis, "Research publication recommendation system based on a hybrid approach", Proc. ACM 20th Pan-Hellenic Conf. Inform., pp. 78-83, 2016.

[36] M. Amami, R. Faiz, F. Stella, G. Pasi, A graph based approach to scientific paper recommendation, in: the International Conference, 2017, pp. 777-782.

[37] C. Wang, DM Blei, Modélisation de sujets collaboratifs pour recommander des articles scientifiques, dans: Conférence internationale ACM SIGKDD sur la découverte des connaissances et lexploration de données, 2011, pp. 448-456.

[38] K. Sugiyama, MY Kan, Exploiting potential citation papers in scholarly paper recommendation, 2013, pp. 153-162.

[39] A. Younus, MA Qureshi, P. Manchanda, C. ORiordan, G. Pasi, Using Microblog Data in a Topic Modeling Framework for Scientific Articles Recommandation, Springer International Publishing, 2014.

[40] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Fast unfolding of communities in large networks, in Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P1000.

Smail Boussaadi Asistant professor, PhD student
Laboratory LIMPAF university of Bouira, Algeria
sboussaadi@gmail.com

Dr Hasina Aliane Hassina Aliane, PHD
Director of Information Sciences R&D Laboratory
Head of Natural Language Processing and Digital Content Team.
Research Center on Scientific and Technical Information, Algiers, Algeria.
haliane@hotmail.com

Prof Ouahabi Abdeldjalil 1.Polytech Tours, Imaging and Brain, INSERM U930, University of Tours, Tours, France
2.University of Bouira, Computer Science Department, LIMPAF, Bouira, Algeria.
ouahabi@univ-tours.fr

Dihia Houari IT Master II university of Bouira Algeria.
diihia.hd@gmail.com

Malika Djoumagh IT Master II university of Bouira Algeria.
djoumaghmika@gmail.com

Multi-Conference OCTA’2019 on:

Organization of Knowledge and Advanced Technologies

Unifying the scientific contributions of the following conferences: SIIE’2019 & ISKO-Maghreb’2019 & CITED’2019 & TBMS’2019 February 6-7-8, 2020 Tunis (Tunisia)

Des mosaïques antiques aux codex islamiques : continuité ou ruptures ?

Khouloud Derouiche

Khaled Kchir

Université de Tunis

2020

Titre :

Des mosaïques antiques aux codex islamiques : continuité ou ruptures ?

Sujet

A partir des mosaïques de la période romaine conservées au Musée du Bardo, nous testerons la pertinence de l'hypothèse de la persistance des influences artistiques antiques à travers de nouveaux supports : les manuscrits, le bois et la céramique. L'examen et l'analyse de certains documents de la période aghlabide et fatimide, ziride ou hafside nous permettront de confirmer ou d'infirmer ces influences.

Le corpus

Mosaïques romaines (Collections du Musée national du Bardo)

Reliures des manuscrits d'époques aghlabide et hafside (VIII-XV^{e} siècles) (La Bibliothèque de Raqqada et la Bibliothèque Nationale de Tunis).

Bois, (d'époque aghlabide VIII-IX^{e} siècles)

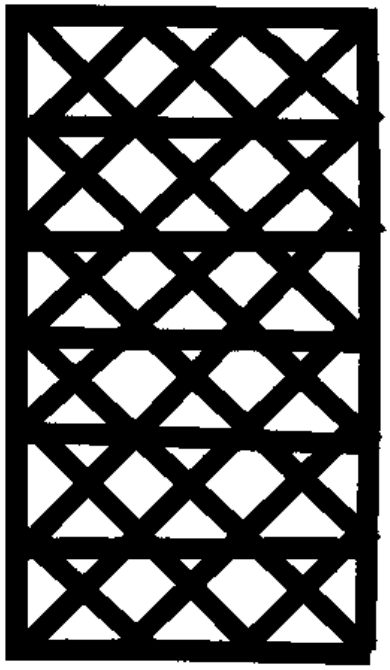
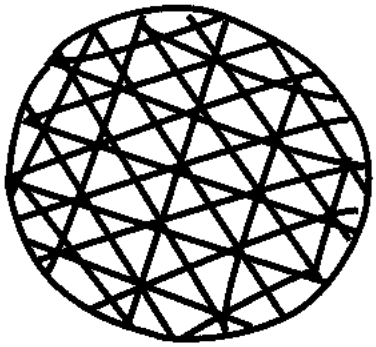
Céramiques et tissus (d'époque fatimide, IX-X^{e} siècles)

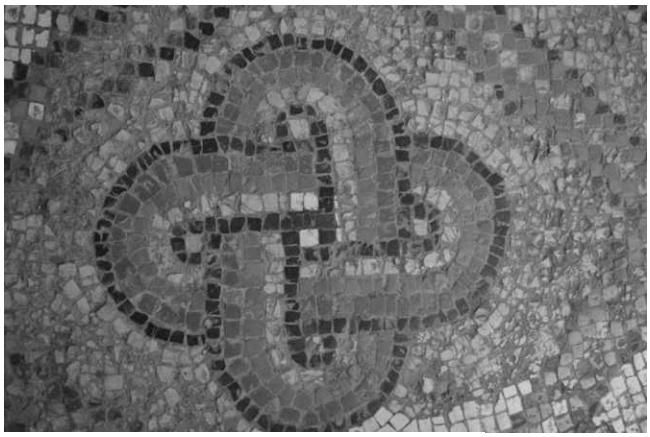
Le Musée du Bardo possède diverses collections représentant le patrimoine depuis la haute antiquité à nos jours. Parmi les plus importantes, les collections de mosaïques romaines tiennent une place de choix. Elles se répartissent dans plusieurs salles du musée. Les tableaux que nous avons sélectionnés sont situés à l'entrée du rez-de-chaussée et au premier étage. Ces mosaïques ont été transférées au musée durant la première moitié du XXème siècle.

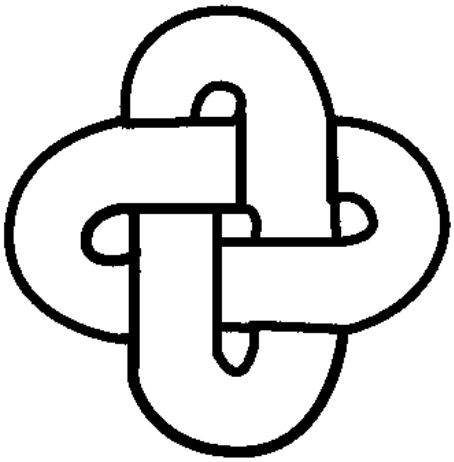
Outre les mosaïques, nous avons sélectionné un nombre d'objets reflétant la richesse du patrimoine tunisien durant le Moyen âge : Des pièces de céramique d'époque aghlabide (provenant des fouilles de Sabra al-Mansuriyya) ainsi que des pièces de céramique hafside. Un panneau en bois provenant de la Grande Mosquée de Kairouan et un morceau de tissu fatimide.

Tous ces panneaux et objets exposés au Musée du Bardo, ont été sélectionnés pour nous permettre d'effectuer des comparaisons avec les motifs décoratifs des manuscrits (Ms). A cet effet nous avons établi un corpus de manuscrits dont les reliures sont décorées. Ces supports relativement fragiles, laissent encore voir la richesse et la variété des références artistiques des relieurs dans les ateliers de Kairouan ou de Tunis. Ces reliures proviennent du fonds de la bibliothèque de Kairouan, (datés de la période aghlabide et de la Bibliothèque nationale de Tunis (Deux spécimens d'époque hafside)

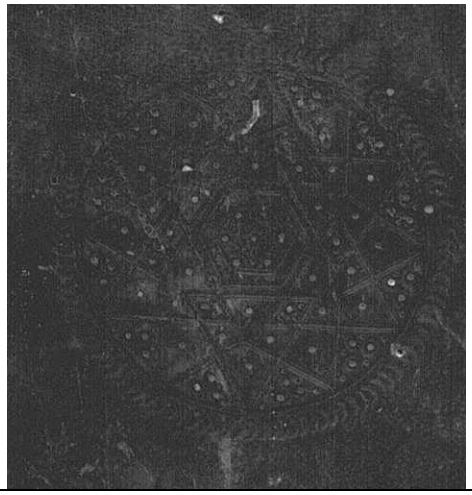
Par ailleurs, nous signalerons les influences entre artisans de domaines différents aux époques aghlabide, fatimide ou hafside, à partir des matrices antiques que nous avons évoquées.

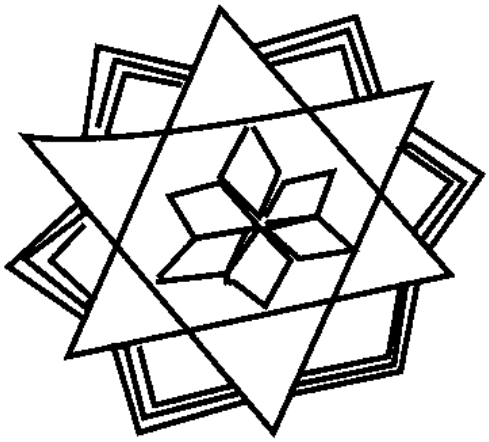
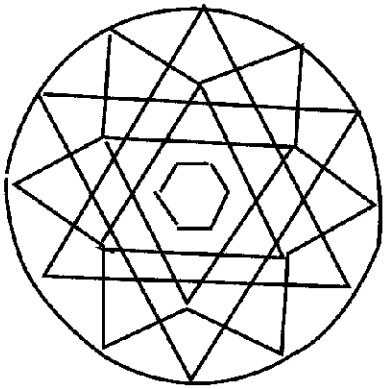
 	  
Raqqada Ms. n° : 064	Le Bardo, Céramique hafside (14e – 16e siècles J.C) Influence hispano-andalouse
Des lignes obliques et horizontales composent des carrés et des rhombes	

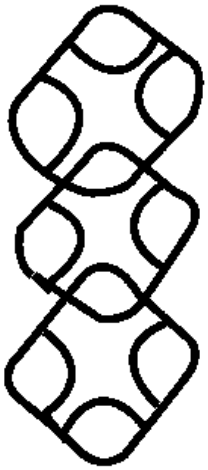
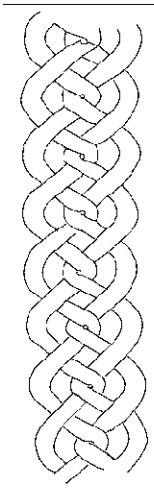


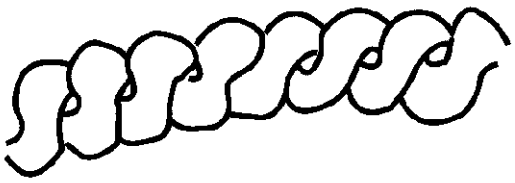


Bardo : Rez-de-chaussée (à l'entrée)	BNT Ms. n° : 6639
Deux formes ovales imbriquées incrustées dans un panneau de mosaïque, ont été reprises par le relieur du Ms 6639 de la BNT.	

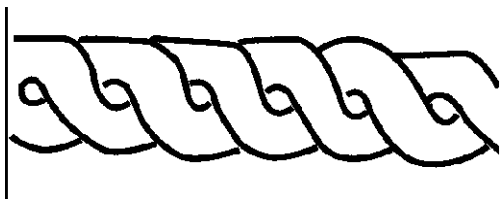


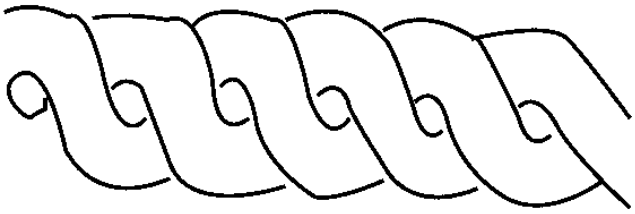
	
Le Bardo : Céramique aghlabide (IXe siècle)	BNT n° :7650
Décors géométriques : 01- Plat a marli, large base (22.8 cm) décor vert et brun sur fond jaune. Une rosace formée de six losanges dessinée au centre d'une autre rosace de six losanges Ce décor géométrique produit un effet de profondeur. 02- Même motif décoratif est repris sur la reliure du Ms 7650.	

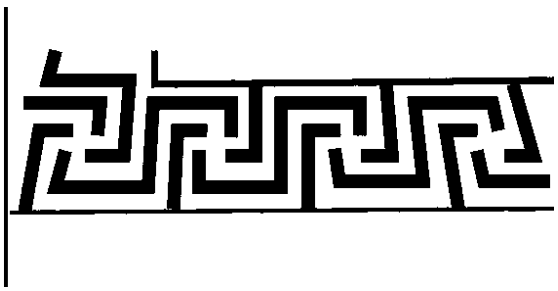
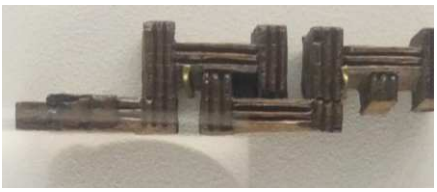
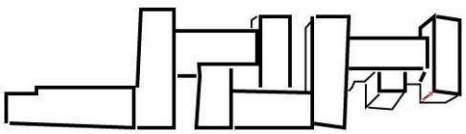
 	 
Raqqada Ms n°: 041	Le Bardo : période fatimide Egypte
La médaille du manuscrit 041 et le tissu fatimide est décorée d'un rhombe imbriqué	



Le Bardo : céramique Sabra al-Mansûriya IVeH/Xe ap JC	Raqqada Ms n°: 078
Differents supports bordés par la lettre S latine (une reliure et céramique aghlabide.)	



 	
Bardo : Rez de chaussée	Raqqada n° : 087
Deux supports différents bordés par une tresse	

 	 
Le Bardo :	Kairouan : Grande mosquée

Décors géométriques imbriqués sur le bois du minbar de la grande mosquée de Kairouan rappelle clairement la mosaïque romaine.

Utilisation des hachures dans tous les supports (dans plusieurs ornements)	

Le centre des manuscrits de Raqqada, renferme un trésor de 5 000 manuscrits précieux et uniques au monde. Cette collection provient de la Bibliothèque de la Grande Mosquée de Kairouan.[1] Riches et variés de par leurs supports, leurs reliures, leurs écritures, leurs calligraphies, ces manuscrits sont aussi différents par leurs contenus thématiques (Coran, Hadîth, théologie, littérature, mathématiques, botanique, médecine, etc …)

Kairouan au cours des X et XI siècles était le centre le plus important de la fabrication du livre qui exportait vers l'Égypte, la Syrie et l'Espagne.

Lors d'une visite au Musée du Bardo, nous avons découvert une similitude entre les motifs romains et aghlabides représentés dans les panneaux de mosaïque. Les manuscrits portent les marques de cette similitude. Ce sont essentiellement des décorations géométriques qui combinent des lignes droites, des courbes et des cercles. Nous avons également remarqué l'utilisation de la technique des hachures, principalement présente dans les manuscrits et dans la céramique.

Ces motifs sont constitués de lignes réfractées et entrelacées, de formes ovales qui se chevauchent, et également d'étoiles hexagonales. Nous retrouvons également des polygones de carrés, de triangles et autres. Et les motifs lyriques constitués de lignes torsadées enroulées ensemble, dont le plus important est le cadre en forme de tresse qui a émergé à l'époque hellénistique. Cette tresse est reprise à l'époque romaine pour cadrer les portrais des princes. Ce même cadre se maintient au cours de la période aghlabide.

[1] Rammeh Mourad, *Manuscrits précieux du patrimoine de Kairouan*, Ministère de la culture, Tunis, 2009.

Nous en déduisons que malgré la diversité des supports à travers les siècles et en dépit des ruptures politiques, qui a contribué à redessiner les frontières, le domaine artistique a su préserver une continuité décelable dans les formes géométriques et les dessins étudiés.

Ces motifs ont peut-être évolué pendant la période islamique. Nous les retrouvons identiques, parfois comme la tresse. D'autres ont été modifiés, de sorte que la forme se développe quelque peu, mais elle conserve sa composition originale.

La différence dans ces décorations est peut être due à la nature des supports et les spécificités des ateliers. Cette similitude et cette harmonie sans frontières géographiques ni historiques, feront l'objet de notre étude.

Bibliographie

شبوح ابراهيم، **سجل قديم لمكتبة جامع القيروان**، مطبعة مصر،1957.

شبوح ابراهيم، **المخطوط**، الوكالة القومية لاحياء واستغلال التراث الاثري والتاريخي، تونس،1989

المحجوبي (عمر)، البلاد التونسية في العهد الروماني، اوتار تبر الزمان، 2016.

Aounallah Samir, « Un monument, un musée, je suis Bardo », Agence de mise en valeur du patrimoine et de promotion culturelle, Institut National du Patrimoine, Tunis, 2016.

Daoulatli Abdelaziz, « De l'Ifriqiya à la Tunisie de 687 à 1881 », dans *Le Bardo la grande histoire de la Tunisie musée sites et monuments*, M'hamed Hassine Fantar, Samir Aounallah, Abdelaziz Daoulatli, Fondation Kamel Lazaar, Tunis, 2015.

Kairouan et sa région, Centre de publication universitaire, Université de Kairouan, Faculté des lettres et sciences humaines, Département d'archéologie, Tunis, 2013.

Rammeh Mourad, *Manuscrits précieux du patrimoine de Kairouan*, Ministère de la culture, Tunis, 2009.

Yakoub Mohamed, L*e musée du Bardo*, Agence nationale du patrimoine, 1996.

Zbiss Slimane Mostfa , *L' art musulman en Tunisie, depuis l'avènement des Aghlabides jusqu'à l'avènement des Almohades (184 – 555 H / 800 – 1160 J.C)*, 3éme série, volume 2 , INP, Tunis 1978.

Avatar Technology for the educational support of deaf learners: A Review

Y., Bouzid, and M., Jemni, *Research Laboratory LaTICE, University of Tunis, Tunisia*

Abstract— This paper addresses a particular view to the importance of adopting avatar technology within educational contexts as it could play an important and significant role in helping students with hearing disabilities overcome the academic difficulties that they face and supporting them to develop their learning skills. We discuss in the first part of this work the important factors that may affect the deaf education. In the second part, we explore and analyse the potential benefits of using of avatars and virtual environments in supporting the achievement of these students and their learning goals. The third part of this paper provides a brief overview of several virtual humans technologies used.

Index Terms—Avatar Technology, E-Learning, Technology-Assisted Learning, Learners with disabilities

1 Introduction

According to the Convention on the Rights of the Child, all children have the right to education that meets their needs and guarantees them full integration in their society regardless of their individual differences or difficulties. Even so, ensuring a full and equal access to a quality education is still an ongoing concern for the majority of deaf communities' members all around the world. Unfortunately, a great number of students with hearing disabilities placed in general education classes, and having access to the general education curriculum, remain unable to achieve results commensurate with their abilities or at levels equivalent to those of their peers without disabilities. For this reason, an increasing attention has been drawn towards the design and development of accessible computer applications for these individuals.

Actually, the usage of assistive technology (AT) is vital in preparing the deaf and hearing impaired students with appropriate learning environments [1]. With assistive technology, Deaf and Hearing Impaired students (DHI) could access curriculum materials in their own suitable way, participate independently in classroom activities alongside their peers and achieve their full potential. To date, an extensive range of innovative devices and services has been designed to meet these goals, it can be broadly classified into three general categories: hearing technologies (e.g., Personal amplification devices); alerting products (e.g., Visual or vibrating systems); and communication supports that offer alternative solutions to access information through vision (e.g., Real-time captioning, video rely interpreting, video conferencing, signing avatars and interactive learning resources) [2].

Within these different types of assistive technologies, the use of 3D animated avatars has been identified as a successful practice that can support the teaching of literacy to students who are deaf or hard of hearing [3], [4], [5]. These digital agents have the potential to act as a powerful communication medium for learners to display knowledge in their first language and make instructional materials completely accessible to them. As mentioned by Vesel [6], the content knowledge of students who are deaf or hard of hearing improved when they used a digital science program that included a signing avatar. Furthermore, by appearing on screen as embodied entities, whether humans, or anthropomorphized characters and animals, these agents can increase effectively learners' attention and motivate them to keep interacting with the content presented. In this sense, Wang et al. [7] pointed out that, from a preliminary assessment of usability, avatars seem to have a beneficial effect on learner motivation and concentration during learning.

Nowadays, software for creating signing avatars has been developed and applied, in an increasing number of countries, to generate three-dimensional representations of sign language communication to deaf learners. Signing avatars can be adopted now for rendering stories, poems, scientific terms, or any written text into signed languages. Thanks to its flexibility and cost-effectiveness, this exceptional technology has opened up new vistas for communication and knowledge acquisition for DHI learners in ways that were unimaginable even a few short years ago.

In this paper, we will pay special attention to the importance of use avatar technology within educational contexts for students with hearing disabilities given their significant role which can play in helping them overcome the academic difficulties that they face and supporting them to develop their learning skills as well. We will discuss in the first part the important factors that affect the deaf education; secondly we will explore and discuss the benefits for the use of avatars and virtual environments in supporting the achievement of DHI. The third part of this paper reviews some virtual humans technologies commonly used for educational purposes.

- *Yosra Bouzid is with the Research Laboratory of Technologies of Information and Communication & Electrical Engineering (LaTICE), Tunis, E-mail: yosrabouzid@hotmail.fr*
- *Mohamed Jemni is with ICT Department at the Arab League Educational, Cultural and Scientific Organization. E-mail: Mohamed.jemni@alecso.vc.*

2 Factors Affecting Deaf Education

Illiteracy and semi-literacy are seriuous problems among deaf people. Clearly, there are a number of factors that might adversely affect the academic performance and progress of these members. Some factors are at student level, some at institutional or programme level and others at structural level. We will try to examine, in this section, those factors under the light of literature review and case studies.

2.1 Mode of Communication

The mode of communication is the most critical factor determining deaf pupils' attainment [8], [9] despite some evidence that is not a major influence [10]. Some researchers view that it is very important that communication between teachers and deaf pupils work properly to assimilate an educational program. If it does not, the pupil will surely have difficulties taking in the provided education [11]. Such problem could happen, typically, if teachers and pupils do not have the same standard in sign language. Unfortunately, it is interesting to point out here that most teachers have a standard of sign language that is below that of deaf learners who have sign language as their first language. Most importantly, the majority of these instructors have little or no exposure to educating children with any degree of hearing loss; they feel frequently ill-prepared to meet the needs of these signers [12], [13]. The issue of training of the teachers of hearing-impaired students to become skilful signers still exists, especially in developing countries [14].

Typically, the use of a sign language interpreter can help overcome this communication gap by interpreting the teacher's explanations and requests to the student and by providing feedback or questions to the teacher from the student if the student does not have sufficient expressive language. Nonetheless, this method is at best impractical due to the lack of availability of a specialized sign interpreter. In fact, there is a continuing shortage of qualified interpreters. For example, in public school in USA, students in STEM (Science, Technology, Engineering and Mathematics) courses frequently receive content translation from either unqualified interpreters, interpreters unfamiliar with concepts of the discipline, or unfamiliar with necessary specialized vocabulary [15]. Moreover, in such mediated instruction, DHI students face further challenges as they are required to divide their visual attention between the instructor, the interpreter, and any visual display [16].

2.2 Language Acquision

Another important factor connected to the low level of educational attainment among deaf learners is the poor knowledge of the dominant language. It is necessary to understand that the majority of deaf children are born to hearing parents with little or no experience of deafness or knowledge of how to communicate with a deaf person. These parents often communicate with their children through home-made signs, pointing and gesturing. For this reason, deaf children arrive at school without a language [17], [18]. Secondly, in the children's prime years of learning, they miss out on fairy tales, story book reading, and the language structure of their parent's language and of their own language. So, due to the parents' inability to communicate with them, these deaf children get no explanations of why they can and cannot do things. More importantly, deaf children miss out on incidental learning and language development that take place daily in conversations around them or through information that would normally be gleaned from playmates, radio, television and other sources [19]. This lack of background knowledge and information puts them at a major disadvantage compared to hearing children starting school at the same age. According to Meadow [20] for children born with a profound congenital hearing loss, the fundamental deprivation is not that of sound, but of language.

2.3 Instructional programs and approaches for teaching deaf learners

Other factors to consider, in deaf education, are instructional programs and approaches used in teaching learning skills for DHI students. According to its 2011 report about deaf people and human rights, the World Federation of Deaf (WFD) note that the majority of current deaf education programs and policies do not respect the linguistic human rights of these signers. Indeed, most deaf education programs fall into the language deprivation category described in theoretical models of education of linguistic minorities. Language deprivation for deaf members means ignoring the use of sign language as a basic communication means, as a language of instruction and as a school subject (WFD, 2011). Undoubtedly, such fact has serious implications for their educational attainment and cognitive development.

3 Learning Through Avatars

As we have seen in the previous section, there are many factors that might influence the DHI students' performance.Interestingly, the common denominator of these learners' failures is not their limited mental abilities, as many assume, but their inability to rely on their first language for accessing learning content. For this reason, sign language and visual strategies should be made available to them to accommodate their needs and maximize their academic potential. Nowadays, technology developments clearly have a very important role to play in achieving these goals since they can give rise to new teaching and learning facilities. Debevc et al. [21] believe that the use of technology within educational contexts has become inevitable for students with disabilities since it can be an assistive tool providing the support needed to accomplish a task. In turn, Kuzu [22] argues that a properly designed e-learning environment could motivate students and gets them to take an active part in the learning process.

For deaf and hearing impaired students, educational technologies involving sign language, such as computer-animated avatars, can offer a promising medium for more

personalized learning adapted to their unique needs. Signing avatars are animated characters that communicate in sign language through hand gestures, body movements and facial expressions. They can be generated by a computer based on spoken audio or written text as input. These graphical entities are flexibly configurable in terms of appearance and can principally be produced anywhere at any time with a relatively low cost, so that they can easily be embedded in any educational application to deliver sign language utterances. Actually, rendering sign language, in the form of 3D avatar animations, holds considerable potential for improving learning experiences and outcomes of young signers, whether through displaying knowledge in dynamic visual form, or through affording the educators and parents the opportunities to construct and present signing linguistic educational material in the classroom and beyond. The following sections discuss more the key benefits and advantages provided by animated avatars in learning.

3.1 Pedagogical Advantages

There is much evidence citing the pedagogical benefits of using virtual environments and animated avatars for supporting DHI students' achievement. Perhaps the most important benefit is their role to give these students the opportunity to master content in a way that meets their needs and practice literacy skills more effectively. Indeed, many studies have explored the potential of avatar environments to act as powerful communication media for displaying knowledge and understanding to DHI students [23]. Findings indicated that such technology could be used to increase students' comprehension in the content through a purposeful and well-focused communication. One of the studies referenced, conducted by Lang and Steely [24], found that when science content was presented in a short text screen with a corresponding signing animation explicating the text passage significantly greater knowledge gains resulted for students who are deaf than in traditional classroom experiences that did not include this triad.

Another apparent benefit of inclusion avatar technology in e-learning platforms is its ability to inspire and engage students to take a more active role in learning [25]. The mere presence of a lifelike agent may increase the student's attention, arousal and interest to perform the task well [26]. It is important to understand that motivation and engagement pose a significant challenge for students with special needs who frequently find their impairments interfere with their ability to engage in instructional activities, which further undermines their motivation. For that, the employment of 3D animated characters can be a great way to actively engage these students in the learning process in order to perceive instruction as relevant through its use. This claim is further supported by Antonacci et al. [27] who reported that engagement in virtual environments enables students to experience learning opportunities that would not normally have been easily accessible. Falloon [28] argued that enhancing learner engagement can support the development of higher level thinking capabilities such as interpretation, analysis, evaluation and problem-solving.

3.2 Technological Advantages

When compared to other signing mediums like digitized videos of real signers, virtual signing avatars shows many technological advantages. As is well known, the process of creating signing videos is expensive and inefficient with difficulties in reproduction: (a) good quality video needs relatively sophisticated recording facilities as it is necessary to use specific equipments and trained people who deeply know the sign language (b) ensuring continuity is problematic with video when materials are updated since the same signer, clothing, lighting, background and camera settings must be maintained, (c) stitching sequences together is impractical, requiring a complete re-shoot if minor changes are needed or versions in different dialects are required, (d) transmitting and storing high quality video consumes significant resources and introduces delays. These problems can be easily avoided by using avatar technology because: (a) virtual avatars can generate real time content, so continuity is not a problem since details of the content can be edited at any time without having to rerecord whole sequences, (b) signed content can be consistently updated, viewed from different angles and at different speeds, and performed by virtual characters of different ages, ethnicities, and genders, (c) parts of a sequence can be changed, or selected on the basis of dialect, without the need to re-create unaffected parts of a sequence, (d) the transmission or storage demands of the signing animation are negligible compared with the corresponding video. Certainly, these aspects may make signing avatars preferable to videos [29].

4 Related Work

The research in deaf education using avatar has diversified over the last years. Several projects have been devoted to the development of deaf students' literacy in their local vernacular language through the medium of sign language avatars. In this section we outline some of the many projects that use a variety of methods to drive their avatar.

4.1 Signing math dictionary & Signing Science Dictionary

Signing Math Dictionary, a product developed as collaboration between TERC and Vcom3D, is an illustrated, interactive 3D sign language dictionary [30]. It contains 705 math terms defined in both American Sign Language (ASL) and Signed English (SE), designed for use by students in grades 4- 8. Science Dictionary (SSD) is an interactive 3D sign language dictionary with 1,300 science terms defined in both American Sign Language (ASL) and Signed English (SE).

Indeed, these two applications use the signing avatar accessibility software to develop an illustrated, interactive, 3D sign language dictionary of science terms and definitions.

4.2 Sign Language Tutor

Sign Language Tutor is an interactive platform for learning Macedonian Sign Language in an easy and entertaining way. It represents a collection of games that ease the learning and increase mental and memory skills of the deaf and hard of hearing children. The central part of this project consists of 3D animations of a girl that signs the chosen alphabet character or some objects [31]. Such systems can be instrumental in assisting sign language education, especially for non-native signers.

4.3 SMILE

SMILE is an immersive learning environment designed to sign mathematics concepts to K-4 DHI students who know ASL. They provide equal access and opportunities by overcoming known deficiencies in science, technology, engineering, and math (STEM) and they offer a teaching model that improves deaf education in general. SMILE is an immersive game in which DHI and hearing children (ages 5-11) learn standards-based math and science concepts and associated ASL signs based on interaction with fantasy 3D characters and objects [32].

4.4 Tawasoul

Tawasoul is a research project conducted by the Computer Sciences Department in King Saud University. It was specifically designed as a multimedia educational tool for teaching Arabic Sign Language (ARSL) hearing impaired children, their parents, and others who are interested in learning ARSL. It consists of three parts: (a) Translator which allows users to enter an Arabic text and view the Arabic signs that are related to the entered text; (b) Dictionary which provides basic vocabulary guide for users who want to learn Arabic Sign Language. It consists of a number of categories; each one contains a related group of words. (c) Finger Spelling: it can be utilized as a sign language editor to help users to write documents in sign alphabetic letters by converting the entered Arabic text to sign language text [33].

4.4 Tunisigner

tunisigner platform [34], [35], [36] is a part of the broader project WebSign [37]. It includes three main components: Sign Language Dictionaries, a Memory Match Game called MemoSign and an application for learning Quran in sign language called Al Bayen.

The set of bilingual dictionaries helps deaf users find the sign language utterance equivalents of a written word or a SignWriting transcription. DHH users could choose, from the sample of SignWriting notations, the term to render and then click and visualize its corresponding interpretation in form of 3D signing animations.

MemoSign is basically a version of the Memory Match Game in which cards are flipped to their backside to not show their content and the player must find a pair of cards from the bunch of cards available that is matched with each other. The present game aims essentially to boost the vocabulary acquisition for deaf children in both signed and spoken written languages.

Fig. 1. Five bilingual sign language dictionaries are provided on tunisigner platform: American Sign Language Dictionary, Egyptian Sign Language Dictionary, Tunisian Sign Language Dictionary, Brazilian Sign Language Dictionary and French Sign Language Dictionary.

In fact, the proposed game is split up into two different categories: the first category is a pair of cards in which the child must find the correct correlation between the word card, which holds written information, and the SignWriting card, which has the same meaning as the written card but in SignWriting. As they watched the avatar signs a SW notation in SL, deaf learners could easily comprehend the meaning of the transcribed gestures [38], [39].

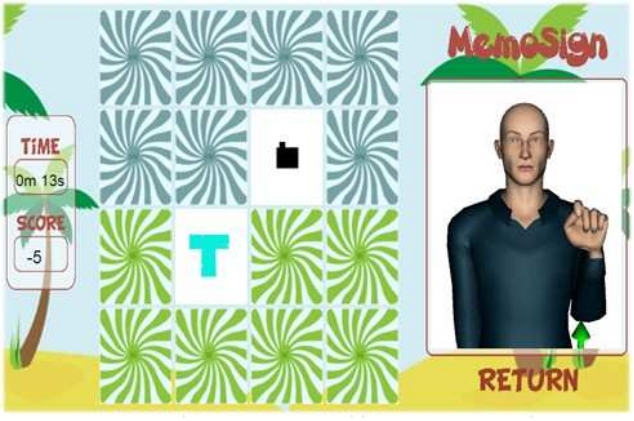


Fig. 2. MemoSign game offers to the learner the opportunity lo learn new vocabularies in sign language using SignWriting notations

Al Bayen application is mainly intended to provide a literal translation of the Quranic verses as well as a written transcription to their meaning in SignWriting. In order to facilitate the deaf users' access to written content in the language they can comprehend and help them to memorize few verses from the Holy Quran, the application provides a visual-gestural interpretation, rendered by a 3D animated avatar, accompanying each SignWriting transcription. Through using Al Bayen

application, the deaf user has the possibility to choose three techniques to read Quranic verses: (a) the first one consists in Finger spelling the letters which compose quran verses, (b) the second one provides sign language utterances equivalents to the verses' meaning by adopting a virtual avatar, (c) and the last one provides a translation of Quran verses in the form of video sequence [40].

Fig. 3. Al Bayen application provides sign language utterances equivalents to the verses' meaning by adopting a virtual avatar.

5 Conclusion

Avatar technology is a fairly new horizon which has opened many educational doors to children, particularly to children with disabilities. Particularly, there seem to be many advantages for learning by using virtual environments and interactive avatars in supporting DHI students' achievement and learning goals.

References

[1] M. Berndsen, and J. Luckner, "*Supporting Students Who Are Deaf or Hard of Hearing in General Education Classrooms: A Washington State Case Study*," *J. Communication Disorders Quarterly*, vol. 33(2), pp. 111-118, 2012

[2] Y. Bouzid, M. A. Khenissi, and M. Jemni, "The effect of avatar technology on sign writing vocabularies acquisition for deaf learners," *Proc. the 16th IEEE International Conference on Advanced Learning Technologies*, pp. 441-447, 2016

[3] N. Adamo-Villani, and K. Hayward, "Virtual immersive and 3d learning spaces", *Emerging technologies and trends* (S. Hai-Jew, Ed.). vol. 2011.

[4] J. Herdich, "Signing Avatar characters for the instruction of K-12 deaf learners." Exploring Instructional and Access Technologies Symposium, New York, NY, 2008

[5] L. C. Wang, and D. Fodness, "Can avatars enhance consumer trust and emotion in online retail sales?," J. International Journal of Electronic Marketing and Retailing, vol. 3(4), pp. 341-362, 2010

[6] J. Vesel, "Signing science", *J. Learning & Leading with Technology*, vol. 32(8), pp. 30-35, 2011

[7] H. Wang, M. Chignell, M. Ishizuka, "Improving the Usability and Effectiveness of Online Learning: How Can Avatars help?," *Proc. Human Factors and Ergonomics Society Annual*, pp. 769-773, 2007

[8] A. Geers, and J. Moog," Factors predictive of the developments of literacy in profoundly hearing-impaired children", *J. Speech and Hearing Disorders*, vol. 52, pp; 84–94, 1989

[9] S. Lewis, "The reading achievement of a group of severely and profoundly hearing-impaired school leavers educated within a natural aural approach", *J. British Association of Teachers of the Deaf*, vol. 20, pp. 1–7, 1996

[10] C. J. Jensema, and R. J. Trybus, "Communicating patterns and educational achievements of hearing impaired students," Washingtom, D.C.: Gallaudet College Office of Demographic Studies, 1978

[11] E. Rydberg, L. Coniavitis Gellerstedt, and B. Danermark, "Toward an equal level of educational attainment between deaf and hearing people in Sweden?" *J. Deaf Studies and Deaf Education*, 2009 doi: 10.1093/deafed/enp001.

[12] M. Doyle, and L. Dye, *Mainstreaming the student who is deaf or hard-of-hearing: a guide for professionals, teachers, and parents*, M.Ed, Parent of hard-of-hearing child, 2002

[13] N. M. Moon, R. L. Todd, D. L. Morton, and E. Ivey, "Accommodating students with disabilities in science, technology, engineering, and mathematics (STEM): Findings from research and practice for middle grades through university education", Atlanta: Center for Assistive Technology and Environmental Access, College of Architecture, Georgia Institute of Technology. 2012, Retrieved from:http://www.catea.gatech.edu/scitrain/accommodating.pdf

[14] G .P. Berent, "*English for deaf students: Assessing and addressing learners' grammar development*, " In D. Janáková (Ed.), International Seminar on Teaching English to Deaf and Hard-of-Hearing Students at Secondary and Tertiary Levels of Education, pp. 124-134, 2001

[15] N. Adamo-Villani, R, and L Wilbur, P. Eccarius, and L. Abe-Harris, "Effects of character geometric model on the perception of sign language animation", *Proc. IEEE 13th International Conference on Information Visualization*, Barcelona, pp. 72-75, 2009

[16] C. L. Baker-Shenk, and D. Cokely, American Sign Language: a teacher's resource text on grammar and culture. Washington, D.C., Clerc Books, Gallaudet University Press, 1991

[17] S. Briggle, "Language and Literacy Development in Children Who Are Deaf or Hearing Impaired". *J. Kappa Delta Pi Record*, vol. 42, pp. 68-71, Kappa Delta Pi, International Honor Society in Education, 2005

[18] M. Marschark, and P. E. Spencer, "Spoken Language Development of Deaf and Hard-of-Hearing Children: Historical and Theoretical Perspectives", In P. E. Spencer & M. Marschark (Eds.), *Perspectives on deafness. Advances in the spoken language development of deaf and hard-of-hearing children*, pp. 3–21, Oxford University Press, 2006

[19] R. E. Mitchell, and M. A. Karchmer, "Demographic and achievement characteristics of deaf and hard of hearing students". In M. Marschark & P. E. Spencer (Eds.), Oxford handbook of deaf studies, language, and education, vol. 1(2), pp. 18-31, New York: Oxford University Press, 2011

[20] K. P. Meadow, Deafness and child development. University of California Press, Berkeley, 1980

[21] M. Debevc, Z. Stjepanovic, and A. Holzinger, "Development and evaluation of an e-learning course for deaf and hard of hearing based on the advanced Adapted Pedagogical Index method," *J. Interactive Learning Environments*, vol. 22(1), pp. 35-50., doi: 10.1080/10494820.2011.641673

[22] A. Kuzu, "The factors that motivate and hinder the students with hearing impairment to use mobile technology," *J. Turkish Online Journal of Educational Technology*, vol. 10(4), pp. 336-348, 2011

[23] K. Melhuish, and G. Falloon, G. "Looking to the future: M-learning with the iPad." *J. Computers in New Zealand Schools*, vol. 22(3), pp. 1-16, 2010

[24] H. G. Lang, and D. Steely, "Web-based science instruction for deaf students: what research says to the teacher", *J. Instructional Science*, vol. 31, pp. 277-298, 2003

[25] S. Deuchar, and C. Nodder, "The impact of avatars and 3D virtual world creation on learning". *Proc. the 16th Annual NACCQ Conference. Palmerston North*, pp. 205-258, 2003

[26] J. C. Lester, S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, "The persona effect: Affective impact of animated pedagogical agents", *Proc. CHI '97*, pp. 359-366. 1997

[27] D. Antonacci, and N. Modaress, "Second Life: The Educational Possibilities of a Massively Multiplayer Virtual World (MMVW)", *Proc. EDUCASE, Southwest Regional Conference*, 2008

[28] G. Falloon, "Using avatars and virtual environments in learning: what they have to offer?", J. British Journal of Educational Technology, vol. 41(1), pp. 108-122, doi: 10.1111/j.1467-8535.2009.00991.x, 2010

[29] P. Lu, "Data-driven sign language animation generation: A survey", (Thesis, City University of New York). Retrieved from http://latlab.cs.qc.cuny.edu/pengfei/lu-literaturesurvey-2011-final, 2011

[30] J. Vesel, and T. Robillard, "Teaching Mathematics Vocabulary with an Interactive Signing Math Dictionar", *J. Research on Technology in Education*, vol. 45(4), pp. 361-389, 2013

[31] N. Ackovska, M. Kostoska, and M. Gjurovski., "Sign Language Tutor – Digital improvement for people who are deaf and hard of hearing," *Proc. ICT Innovations* 2012

[32] N. Adamo-Villani, and K. Wright, "SMILE: an immersive learning game for deaf and hearing children," *In: ACM SIGGRAPH 2007 Educators Program*, pp. 17. ACM, San Diego

[33] A. Al-Nafjan, and Y. Al-Ohali, "A Multimedia System for Learning Arabic Sign Language: Tawasoul," Proc E-Learn: World Conference on ELearning in Corporate, Government, Healthcare, and Higher Education. 2010

[34] Y. Bouzid, and M. Jemni, "An Avatar based approach for automatically interpreting a sign language notation," *Proc. The 13th IEEE International Conference on Advanced Learning Technologies (ICALT)*, pp. 92-94, Jul. 2013, doi:10.1109/ICALT.2013.31

[35] Y. Bouzid, and M. Jemni, "tuniSigner: A Virtual Interpreter to learn SignWriting," *Proc. The 14th IEEE International Conference on Advanced Learning Technologies (ICALT)*, pp. 601-605, 2014, doi:10.1109/ICALT.2014.176

[36] Y. Bouzid, and M. Jemni, "A Virtual Signer to Interpret SignWriting," *Proc. The 14th International Conference on Computer Helping People with Special needs (ICCHP)*, pp 458-465, 2014

[37] O. ElGhoul, and M. Jemni, "WebSign: A system to make and interpret signs using 3D Avatars," *Proc. The Second International Workshop on Sign Language Translation and Avatar Technology (SLTAT)*, Dundee, UK. 2011

[38] Y. Bouzid, M. A. Khenissi, F. Essalmi and M. Jemni, "Using Educational Games for Sign Language Learning - A SignWriting Learning Game: Case Study," *J. Educational Technology & Society*, vol. 19 (1), pp. 129–141, 2016

[39] Y. Bouzid, M. A. Khenissi, F. Essalmi and M. Jemni, "A learning game for deaf learners, "Proc. *The 15th IEEE International Conference on Advanced Learning Technologies*, pp. 418-422, Sep. 2015, doi. 10.1109/ICALT.2015.98

[40] Y. Bouzid, and M. Jemni, "ICT-based applications to support the learning of written signed language," *Proc. Sixth International Conference on Information and Communication Technology and Accessibility (ICTA)*, Dec. 2017

Yosra Bouzid is a Doctor of Computer Science and an instructor at the National Higher School of Engineers of Tunis-ENSIT. She is also a research member at the laboratory of Technologies of Information and Communication & Electrical Engineering (LaTICE). She has published several academic papers in international referred journals and conferences. She has also served as a local organizing and program committee member in various international conferences and as a reviewer in several refereed journals. Her research interests include assitive technology for people with special needs, game-based learning, sign language processing and computer graphics.

Mohamed Jemni. Mohamed JEMNI is a Professor of Computer Science and Educational Technologies at the University of Tunis, Tunisia. He is the Director of ICT Department at The Arab League Educational, Cultural and Scientific Organization from October 2013. He is currently leading several projects related to the promotion of effective use of ICT in education in the Arab world, namely, OER, MOOCs, cloud computing and the strategic project ALECSP-APPS. He produced two patents and published more than 300 papers in international journals, conferences and books. He produced many studies for international organizations such as UNESCO and ITU.

Gamified and self-determined learning environment to enhance students' performance

Mouna Denden[1], Ahmed Tlili[2], Fathi Essalmi[1], Mohamed Jemni[1]
[1]Research Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE), Tunis national higher school of engineering (ENSIT), University of Tunis, Tunisia
[2]Open Educational Resources Laboratory, Smart Learning Institute of Beijing Normal University, China

mouna.denden91@gmail.com
ahmed.tlili23@yahoo.com
fathi.essalmi@isg.rnu.tn
mohamed.jemni@fst.rnu.tn

***Abstract*— Gamification is a prominent approach to foster motivation and further enhance learning performance. However, it is recognized that designing an effective gamified environment is a challenging process. Many researchers, on the other hand highlighted the importance of satisfying students' psychological needs, reported in the self-determination theory, in order to enhance their performance. Therefore, this study presents an ongoing project which is a gamified learning environment based on the self-determination theory, to motivate students and further enhance their performance. Particularly, the proposed approach for designing this environment is based on the satisfaction of students' psychological needs using game design elements .**

Keywords- learning performance; self-determination theory; gamification; game design elements;

I. INTRODUCTION

Video games have proved to be an efficient tool for teaching all age groups and genders for many years. Without a doubt, video games have the potential to provide a high level of motivation [1]. Given this potential, researchers and practitioners think to use the motivational power of video games to motivate users in non-game environment. Specifically, this approach was named gamification. The term "gamification" was introduced in the early 2010s as the use of game design elements such as points, badges and leaderboards in non-game context [2]. By referring to this definition, Werbach [3] stated that each added game design element should have a playful intention in order to be called gamification. For instance, the use of badges in codeacademy educational site evoke motivation and engagement like in game experience. Gamification was applied in many areas such as marketing, health and education with the promise of fostering users' motivation, engagement and promoting changes in behaviors [4]. Particularly, while many studies highlighted the positive effects of gamification in education [2, 4], some others highlighted a negative effect. This was explained by the bad gamification design, according to many researchers [11].

Several game design elements were proposed in the literature to gamify an environment and providing an effective gamification is not a simple process [5]. Specifically, it is found that adding all the game design elements at the same time does not always lead to an effective gamification, which motivate students and further enhance their learning performance [5]. Additionally, there is a lack of theoretical foundation to explain the motivational effects of game design elements. Therefore, motivational theories must be applied in order to explain the motivational process, which is can lead to students' academic success [6]. Particularly, Self Determination Theory (SDT) is one of the most known and successfully applied theories in games [7]. Furthermore, SDT includes a wide range of motivational mechanisms that overlap in part with many of the other perspectives. Therefore this paper presents, in the next section, a new approach to gamify a learning environment based on the SDT, in order to increase students' motivation and further enhance their performance.

II. THE PROPOSED APPROACH

The proposed approach for designing gamified learning environment that enhance students' learning performance is based on the SDT. In particular, SDT contains three basic psychological needs namely competence, autonomy and social relatedness [8], which are described as follows:

- The need for competence refers to the feeling of skillfulness while interacting with an environment.
- The need for autonomy refers to the feeling of freedom while taking decisions.
- The need for relatedness refers to the feeling of being part of a group.

The satisfaction of the three psychological needs (competence, autonomy and relatedness) reported in SDT can facilitate students' internalization of intrinsic motivation which further facilitate students' engagement and enhance their performance [9]. Therefore, our goal is to find the suitable game design elements that satisfy students' psychological needs.

Various game design elements have been used to gamify learning environments. Particularly, to fulfill students' basic psychological needs reported in SDT (competence, autonomy and relatedness) and increase their motivation, eight game design elements were proposed [10]:

- Points are numerical representation that shows students' performance in certain activities within the gamified environment.
- Badges are visual rewards for each achieved goal.
- Leaderboard is a board that shows the ranking of students based on their scores.
- Progress bar provides information regarding students' progress toward a goal.
- Feedback presents teachers' feedback regarding students' performance.
- Chat is instantaneous discussion.
- Avatar is visual presentation a student can choose or upload to his/her profile.
- Levels moderate the level of difficulty based on students' expertise.

Each proposed game design element was choosing carefully in order to fulfill one of the three psychological needs reported in SDT. For instance, badge game design element can show the student's performance while learning. This can make him/her feel competence. Furthermore, the chat game design element can be matched the need for relatedness since it provides a social support for students while learning.

III. CONCLUSION

This study presents an approach for designing a gamified learning environment based on SDT to enhance students' performance. Future work could focus in: (1) validating the effectiveness of the proposed game design elements in fulfilling students' psychological needs, and (2) investigating the efficiency of the proposed approach in enhancing students' learning performance.

REFERENCES

[1] J. Hense, & H. Mandl, "Learning in or with games? Quality criteria for digital learning games from the perspectives of learning, emotion, and motivation theory". In D. G. Sampson, D. Ifenthaler, J. M. Spector, & P. Isaias (Eds.), Digital systems for open access to formal and informal learning (pp. 181-193), 2014. Piraus: Springer.

[2] S. Deterding, D. Dixon, R. Khaled, and L. Nacke, "From game design elements to gamefulness: Defining gamification," Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments, pp. 9–15. New York, NY, USA: ACM, 2011.

[3] K. Werbach, "(Re)defining gamification: A process approach." In A. Spagnolli, L. Chittaro, & L. Gamberini (Eds.), Persuasive technology, pp. 266-272, 2014. Springer, Cham.

[4] M. Denden, A. Tlili, F. Essalmi, and M. Jemni, "Does personality affect students' perceived preferences for game elements in gamified learning environments?," in Procedings of the 18th International Conference on Advanced Learning Technologies (ICALT), 2018, pp. 111-115. IEEE.

[5] Z. Fitz-Walter, D. Johnson, P. Wyeth, D. Tjondronegoro, & B. Scott-Parker, "Driven to drive? Investigating the effect of gamification on learner driver behavior, perceived motivation and user experience," Computers in Human Behavior, vol. 71, 2017, pp. 586-595.

[6] P. Buckley, and E. Doyle, "Gamification and student motivation," Interactive Learning Environments, vol 22(6), 2014, pp. 1–14.

[7] M. Sailer, J. U. Hense, S. K. Mayr, and H. Mandl, "How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction," Computers in Human Behavior, vol. 69, 2017, pp. 371-380.

[8] E. X. Deci, and R. M. Ryan, "Intrinsic motivation and self-determination in human behavior." New York: Plenum Press, 1985.

[9] R. M. Ryan, and E. L. Deci, "When rewards compete with nature: The undermining of intrinsic motivation and self-regulation," In C. Sansone & J. M. Harackiewicz (Eds.), Intrinsicand extrinsic motivation: The search for optimal motivation and performance, pp. 14–54. SanDiego, CA: Academic Press, 2000.

[10] K. Werbach, & D. Hunter, "For the win: How game thinking can revolutionize your business." 2012. Philadelphia: Wharton Digital Press.

[11] Toda, A. M., Valle, P. H., & Isotani, S. (2017, March). The dark side of gamification: An overview of negative effects of gamification in education. In Researcher Links Workshop: Higher Education for All (pp. 143-156). Springer, Cham.

Multi-objective Clustering Algorithm with Parallel Games

Dalila Kessira, and Mohand-Tahar Kechadi

Abstract— Data mining and knowledge discovery are two important growing research fields in the last few decades due to abundance of data collected from various sources. In fact, the exponentially growing volumes of generated data urge the development of several mining techniques to feed the needs for automatically derived knowledge. Clustering analysis (finding similar groups of data) is a well-established and a widely used approach in data mining and knowledge discovery. In this paper, we introduce a clustering technique that uses game theory models to tackle multi-objective application problems. The main idea is to exploit specific type of simultaneous move games, called congestion games. Congestion games offer numerous advantages ranging from being succinctly represented to possessing a Nash equilibrium that is reachable in a polynomial-time. The proposed algorithm has three main steps: 1) it starts by identifying the initial players (or the cluster-heads); 2) then, it establishes the initial clusters' composition by constructing the game to play and try to find the equilibrium of the game. The third step consists of merging close clusters to obtain the final clusters. The experiment results show that the proposed clustering approach gives good results and it is very promising in terms of scalability, and performance.

Index Terms— Data mining, data analysis, clustering, game theory, simultaneous-move games, Nash equilibrium.

1 Introduction

DATA science is one of the most growing research fields over the last few years. It refers to an empirical approach that uses the available big amounts of data to provide answers to wide variety of questions and problems that human beings are enable to treat without the intervention of the machine [1]. Among the several methods and techniques used in data science, cluster analysis (clustering) is a well-established and a wide-used technique. Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters[2].

In fact, game theory offers very attractive rigorous mathematical tools for modeling and resolving a wide part of strategic situations, where some agent's gain depends not only of his choice of decision but on the others' strategies also. Thus, game theory is increasingly used in different types of decision making processes, and in the study of very complex situations in many domains.

Game theory has been used, among other techniques, in clustering. Game theory based clustering algorithms are not numerous; however, they all suffer from a big algorithm complexity. So their major drawback is their usefulness for big amounts of data. This is due to the big complexity of finding one of the most important solution concepts, named Nash equilibrium. This important aspect of game theory has been widely treated in several works, since the 80ies of the last century, in the computer science community[3]–[6], under the theme of computational game theory. But unfortunately, intractability results are the most found, while tractability results are really rare. And the complexity of finding solution concepts remains one of the most important drawbacks of game theory while handling big data is an essential need for modern clustering techniques.

There are many types of clustering algorithms based on various methodologies (see [2] [7], [8], for more details).

As mentioned above, there is no shortage on clustering techniques, but the tendency nowadays is the development of scalable and efficient techniques that allow processing big amounts of data in short time. In addition, the mono-objective clustering may not conform to the diversity and the complexity of modern clustering problems. Hence, this paper is about the proposition of a Multi-objective clustering algorithm based on game theory with parallel games.

2 Background

In this section we present the necessary background in order to introduce, in the next section, our new clustering approach and our clustering algorithm named MOCA-SM (Multi-Objective Clustering Algorithm based on Simultaneous-Move games).

- *D. Kessira is with the Laboratory of Medical Informatics (LIMED), Department of Computer Science, University of Bejaia, 06000 Bejaia, Algeria. E-mail: dalila.kessira@univ-bejaia.dz E-mail: dalila.kessira@gmail.com*
- *M-T. Kechadi is with the Insight Center for Data Analytics, Departement of Computer Science, University College of Dublin, Dublin, Ireland. E-mail: tahar.kechadi@ucd.ie*

2.1 Game theory

Game theory (GT) is the set of analytical tools designed to analyze situations where decision-makers (or players) seek their own profits. The basic assumptions that underlie the theory are that decision-makers pursue well defined exogenous objectives (they are rational) and take into account their knowledge or expectations of other decision-makers' behavior (they reason strategically).[9]

In this section we present the essentials of GT.

2.1.1 Games in normal form

In game theory, the term "game" means an abstract mathematical model of a multi-agent decision-making setting.[10]

Formally, it is represented [11] by the tuple (N, A, u) :

- N = {1, …, n} a set of players, indexed by i;
- $S=S_1*\ldots*S_n$, where Si is a finite set of strategies available to player i. Each vector s= (s1, …, sn) ∈ A is called a strategy profile;
- $u= (u_1, \ldots, u_n)$, where $u_i: S\rightarrow R$ is a real-valued utility (or payoff) function for player i.

The normal form, also known as the strategic or matrix form is when we represent a game via an n-dimensional matrix; it is the most familiar representation of strategic interactions in game theory. A game written in this way represent every player's utility for every state of the world, in the special case where states of the world depend only on the players' combined actions. The normal-form representation is arguably the most fundamental in game theory.[11]

In some contexts, "games" are literally games; for example: choosing how to target a soccer penalty kick and how to defend against it, in other contexts, it is just any strategic situation.

2.1.2 Solution concepts and Nash equilibrium

From here on, we will be interested in non-cooperative games, where it is assumed that individual players self-interested, deviate alone from a proposed solution, if it is in their interest, and do not coordinate their moves in groups.

Reasoning about multi-player games is based on solution concepts, principals according to which we identify interesting subsets of the outcomes of a game [11]. In the section below, we present the most powerful solution concepts in game theory named Nash equilibrium.

2.1.3 Simultaneous vs. sequential-move games

Simultaneous-move games [12] are when all players take their decisions at the same time, consequently, they don't know the decisions made by the other players. Whereas in sequential games [11] players take decision by turn.

The big complexity that usually results from the use of game theory is due to the complexity of finding a solution concepts. Despite that finding such solution concepts, especially Nash equilibrium, for sequential-move games is known to be of a big complexity, using simultaneous-move games doesn't guarantee a better complexity for the algorithm, because it is all about the complexity of finding an equilibrium for the game.

2.1.4 Nash equilibrium

To define Nash equilibrium [13], we need to define the notion of best response [11]. Formally, we denote $s_{-i} = (s_1,\ldots, s_{i-1}, s_{i+1},\ldots, s_n)$, a strategy profile s without agent i's strategy. Thus we can write $s = (s_i , s_{-i})$. If the agents other than i (whom we denote −i) were to commit to play s_{-i}, a utility-maximizing agent i would face the problem of determining his best response.

Definition 1 (Best response): Player i's best response to the strategy profile s_{-i} is a mixed strategy $s^*_i \in S_i$ such that $u_i (s^*_i, s_{-i}) \geq u_i (s_i , s_{-i})$ for all strategies $s_i \in S_i$.

The best response is not necessarily unique. Indeed, except in the extreme case in which there is a unique best response that is a pure strategy, the number of best responses is always infinite. When the support of a best response s^* includes two or more actions, the agent must be indifferent among them; otherwise, the agent would prefer to reduce the probability of playing at least one of the actions to zero.

Thus, the notion of best response is not a solution concept; it does not identify an interesting set of outcomes in this general case. However, we can leverage the idea of best response to define the most central notion in non-cooperative game theory, the Nash equilibrium.

Definition 2 (Nash equilibrium): A strategy profile $s = (s_1,\ldots, s_n)$ is a Nash equilibrium if, for all agents i, s_i is a best response to s_{-i} .

Intuitively, a Nash equilibrium is a stable strategy profile: no agent would want to change his strategy if he knew what strategies the other agents were following. We can divide Nash equilibria into two categories, strict and weak, depending on whether or not every agent's strategy constitutes a unique best response to the other agents' strategies.

Nash equilibrium is a simple but powerful principal for reasoning about behavior in general games [13]: even when there is no dominant strategy, we should expect players to use strategies that are best responses to each other [14]. Intuitively, Nash equilibrium is stable strategy profile where no player would want to change his strategy if he knew what strategies the other players were following[11]. This is the most used solution concept in game theory. On the other hand, it is proved that computing Nash equilibrium is hard for a lot of games' types [5], [6], [15], [16].

2.2 Singleton Congestion games with player specific payoff functions

Milchtaich[17] in 1996, introduced a sub-class of the congestion games presented by Rosenthal [18]. Milchtaich introduces games in which the payoff function associated with each primary factor (or resource) is not universal but player-specific. This generalization is accompanied, on the other hand, by assuming these two limiting assumptions: that each player chooses only one primary resource and that the payoff received actually decreases (not necessarily strictly so) with the number of other players selecting the same primary factor. This sub-class of conges-

tion games, while not generally admitting a potential, nevertheless always possess Nash equilibrium in pure strategies.

So, the cost of a resource *e* for player *i*, will not depend only on the number of players that chose the resource, but it will depend also on the player him-self. So, every player *i* has its own cost function $c_{i,e}$ for every resource *e*, hence the definition below.

Definition 6.1: (SCGPSC: Singleton congestion game with player-specific cost [17]) A SCGPSC game is the tuple $(N, E, S, (c_{ie})_{i\in N,\, e\in E})$:

- ✓ N = {1, …, n} a set of players, indexed by i.
- ✓ E = {1, …, m} a set of resources, indexed by e**.**
- ✓ Strategy set S: S1×… ×Sn, where Si is a set of strategies available to agent i where strategy si ∈ Si is a singleton subset of resources, i.e. set with exactly one element.
- ✓ Cost function $c_{ie}(n_e) \in R$ for player i and resource e, where it can be $c_{ie}(n_e) \neq c_{i'e}(n_e)$ if (i, i') ∈N×N , i.e. the cost function depends on the player him-self and on the number of players that select the resource e.

where $n_e = |\{i : e = s_i\}|$, $(n_1, n_2, \ldots, n_m)$ is called the congestion vector corresponding to a strategy $s=(s_1, s_2, \ldots, s_n)$.

This class of congestion games will be used in our approach to model and resolve the clustering problem.

2.3 Nash equilibrium in SCGPSC

Unfortunately, the Rosertnal's theorem (Theorem 5.1) doesn't hold for congestion games with player-specific cost functions. Malchtaich [17] proved that for this class of congestion games doesn't generally admit a potential, so, best response dynamics don't always converge to Nash equilibrium as it may be cyclic.

However, Matchtaich proved that SCGPSC possess always a Nash equilibrium. He used the proof by induction on the number *n* of players, where it is supposed that an instance of the game with n-1 players has a Nash equilibrium and it's proved that the game with n players has a Nash equilibrium.

For the complete proof see [17], a sketch of the proof is provided here [19]:

- ✓ Put player n apart and take a Nash equilibrium sn-1 for players 1, . . . ,n−1
- ✓ Then introduce player n and let it take its best response
- ✓ Only a player with the same strategy as player n may want to deviate
- ✓ $C_{i\ sn}$ is non increasing, so, no one wants to join resource s_n
- ✓ Suppose player i_0 moves from resource s_n to resource e_0
- ✓ Then only a player with strategy e_0, say i_1, may want to move to resource e_1
- ✓ Every player changes its strategy at most once, until a Nash equilibrium is reached.

This existence proof is constructive and implicitly describes an efficient algorithm for finding an equilibrium in a given n-player SCGPSC, by adding one player after the other in at most $\binom{n+1}{2}$ steps [17].

3 Our approach

Our main contribution is the design of a multi-objective clustering algorithm based on game theory with parallel (or simultaneous-move) games. We use the class of non-cooperative simultaneous-move game presented above, named Singleton Congestion game with Player-Specific Cost (SCGPSC) to model the clustering problem as a simultaneous-move game, where players are a subset of the initial dataset, the resources are the rest of the dataset, and the cost function is an optimization function of two contradictory objectives: connectedness and separation.

3.1 Optimization objectives

Our approach is based on optimizing two conflicting optimization objectives, the first is R-Square, the other is the connectivity of the clusters based on the Euclidean distance. Those two objectives where used in [20] a clustering algorithm that uses sequential game theory. R-Square is optimal when the number of clusters is high and the connectivity of clusters is optimal when the number of clusters is low, so the compromise of those two conflicting objectives guarantees a good quality clustering according to [20].

R-square is given by the formula:

$$R^2(C) = \frac{I_R(C)}{I_A(C)+I_R(C)} \qquad (1)$$

Where $I_R(C)$ is inter-cluster inertia which measures the separation of the clusters:

$$I_R(C) = \frac{1}{n}\sum_{i=1}^{K} |C_i| * d(ch_i, g), \qquad (2)$$

Where ch_i is the cluster-head of cluster C_i, g= (g_1, ... , g_j, . . . , g_m) and g_j is the gravity center of the dataset along the j^{th} dimension.

$$g_j = \frac{1}{n}\sum_{i=1}^{n} o_{ij}. \qquad (3)$$

$I_A(C)$ is the intra-cluster inertia which should be as weak as possible in order to have a set of homogeneous clusters. It is given as:

$$I_A(C) = \frac{1}{n}\sum_{C_i\in C}\sum_{\omega\in C_i} d(\omega, ch_i). \qquad (4)$$

Where: ch_i is the cluster-head of the cluster C_i.

The connectivity measure is intended to assign similar data into the same cluster. It is given by the formula:

$$Connc(C) = \sum_{i=1}^{K} \frac{|C_i|}{n} * Connc(C_i), \qquad (5)$$

Where $Connc(C_i)$ is the connectivity of the cluster C_i, and it is given by the formula:

$$Connc(C_i) = \frac{1}{|C_i|}\sum_{h=1}^{|C_i|} \frac{\sum_{j=1}^{L} \chi_{h,nn_{hj}}}{L}, \qquad (6)$$

Where :

$$\chi_{r,s} = \begin{cases} 1, & if\ r,s \in C_i \\ 0, & otherwise \end{cases}$$

And nn_{hj} is the j^{th} nearest neighbor of object h and L is a

parameter indicating the number of neighbors that contribute to the connectivity measure. The connectivity value should be maximized.

Heloulou et al. [20] have defined the product $\varphi(C)$ that combined the two objectives and should be maximized:

$$\varphi(C) = R^2(C) * Connc(C) \qquad (7)$$

3.2 Model

The game to play is a single-act nonzero-sum multi-player singleton congestion game with player-specific cost function.

A game is a single-act game if every player acts only once; otherwise the game is multi-act. Nonzero-sum game is when there is no net gain or net loss, and the gain of a player doesn't mean the loss of the other. Multi-player game is when more than two players participate in the game; otherwise it is two-player game. Singleton congestion game is when every player chooses a set of one and only one resource. And the cost function of a congestion game is player-specific when it depends on both the number of players choosing this resource and on the player himself.

The players of this game are the objects with more density around them.

Hence, the game G is defined by the tuple:

$$\boldsymbol{G = (N, E, (X_k)_{k\in N}, (co_{ek})_{e\in E, k\in N})} \qquad \textbf{(8)}$$

Where:

$$\boldsymbol{N} = \{\ ch_i \mid ch_i = \underset{k:o_k\in O_i}{argmin}\, dm(o_k),$$

$$O_i = O_{i-1}\backslash\{ch_{i-1}, o_j \mid dis[j][ch_{i-1}] < \frac{Dmax}{n_0}\},$$

$$i = 0..n_0 - 1,\ j = 0..|O_{i-1}|\ \}$$

Where: dm is the dissimilarity of an object, $Dmax$ is the maximum distance between two objects in the dataset, and n_0 is the initial number of players.

$$\boldsymbol{E} = \{o : o \in \mathrm{O}\backslash\mathrm{N}\}$$

$$\boldsymbol{X_k} = \{\{o\} \mid o \in E\}, k = 0..|\mathrm{N}| - 1;$$

$$\boldsymbol{co_{ek}(n_e)} = \begin{cases} -\left(Connec\left(C_k^{(t)}\right) * R^2\left(C^{(t)}\right)\right), k = 0..|\mathrm{N}| - 1, if\ n_e = 1 \\ \infty \quad , if\ n_e > 1 \end{cases}$$

$$, \ \mathrm{e} \in \mathrm{E},\ n_e = |\{z: e = x_z\}|$$

And the payoff to player k for strategy profile x is just the negated cost:

$$u_k(x) = -co_k(x)$$

After playing this game, the next step should be the computation of an equilibrium of this game by using one of many solution concepts that exist in game theory. Accordingly, after the resolution of the game we should get an equilibrium allocation of all the conflict-objects to their clusters.

3.3 Algorithms

In this section, we explain in detail our approach of clustering and we present our new algorithm of multi-objective clustering, named MOCA-SM.

MOCA-SM consists of three main steps; the first step consists on the identification of the players of the game played in the second step. Based on dissimilarity measure, objects of the dataset with the less dissimilarity are designed to be cluster-heads, consequently, they will play the game constructed in each iteration of the step2 in order to gain the objects left in the dataset.

The second step is the identification of the composition of the initial clusters. It starts by identifying the set of strategies- which is the set of singleton subsets of the left objects in the dataset, and the identification of the cost of each strategy for each player identified in the first step, and thereafter, playing the SCGPSC over the set of available strategies and reaching Nash equilibrium to determine which object will be affected to which cluster. This step is repeated until there are no more resources to allocate or no player wants to play anymore, i.e. all players decide to stop playing because they don't gain any additional gain if they continue to play.

The third step is the elaboration of final clusters by merging the clusters resulted from the precedent step. The merging is based on dissimilarity measure where the most similar clusters are merged until a stopping criterion is satisfied. The result of this step is the final clusters. Our approach is described in Algorithm1.

Algorithm 1

Inputs: Dataset O=$\{o_0, o_1, \ldots, o_{m-1}\}$
Outputs: set of clusters

1: t←0
2: compute distance matrix dis between all objects of O
3: **Algorithm 2:** Identification of initial players, the set ch
4: **repeat**
5: t←t+1
6: **Algorithm 3:** construct the game in formula (8)
7: **Algorithm 4:** compute Nash equilibrium
8: **for** each player
9: **if** Nash equilibrium enhance the clustering **then**
10: Allocate the Nash equilibrium strategy for player
11: **else**
12: Player out of game
13: **end if**
14: **end for**
15: **until** (O is empty or all players are out of game)
16: **Algorithm 5:** Merge resulted clusters
17: assign the left objects to the closest cluster

Step1: Identification of the initial players

This step aims at the identification of the number of the initial players, and also the identification of the players themselves. Players are objects from the dataset which have the most density around them; they will be cluster-heads during the clustering. The initial number of clusters is estimated using the formula

$$n_0 = m/L \qquad (9)$$

Where: m is the size of the dataset; and L is the number of neighbors to use to find the connectivity of each cluster.

After that, the dissimilarity of each object compared to all other objects in the dataset is computed using this formula:

$$dm(o) = \frac{1}{m-1}\sum_{oi \in O, oi \neq o} dis[o][o_i] \qquad (10)$$

Objects with the smallest value of *dm* are the objects with the most density around them; consequently, they are chosen to be cluster-heads to play over the other objects. After each time a cluster-head is chosen, the objects around it, based on the dissimilarity, are deleted in order to have cluster-heads as disperse as possible. This operation is repeated until there is no objects left or until the number of the chosen cluster-heads reaches n_0.

Algorithm 2 : Identification of initial players

Inputs: Dataset O={$o_0, o_1, \dots, o_{m-1}$}
Outputs: Set ch of initial cluster-heads (players) ch={$ch_1, ch_2, \dots, ch_k$ }

1: compute n0, the initial number of players using formula (9)
2: compute dissimilarity of each object oi of O : dm(oi) using formula (10)
3: O′← O
4: repeat
5: c ← $\text{argmin}_{i:o_i \in O}$, dm ($o_i$)
6: ch.add(oc)
7: O′\ oc
8: O′\ $\left\{o_j \in O' \mid D[c][j] < \frac{Dmax}{n0}\right\}$
9: **until** (O′ is empty or |ch| = n0)

Step2: establishment of initial clusters' composition

This step aims at determining the composition of initial clusters by constructing the game described in formula (8). The players of the game are the objects with the highest density around them. This game is played several times until all objects are allocated or when no player wants to play again.

Algorithm 3 presents the Identification of initial players, it takes in input the dataset and determines the set of initial cluster (players) C={$c_1, c_2, \dots, c_k$ }

Algorithm 3 : Formulation of the game

Inputs: players ch={$ch_1, ch_2, \dots, ch_k$ }, dataset O={$o_0, o_1, \dots, o_l$},
Outputs: Matrix *Cost* containing the costs of all strategies to all players

1: Identify available strategies over the set O
2: **for each** player ∈ ch
3: **for each** strategy
4: **for each** value of the congestion vector
5: compute the cost of the singleton strategy for the player using formula
6: assign the cost to the matrix Cost
7: end for
8: end for
9: **end for**

Step 3: Merging close clusters

After having n_0 initial clusters from step2, this step is about merging them until we have the exact number of clusters. We use the method of determining the number of clusters based on the improving of the silhouette index. First, the distance between each pair of the initial clusters is calculated using the formula:

$$disC\left(c_i, c_j\right) = \frac{1}{|c_i|}\sum_{oi \in ci}\left(\frac{1}{|c_j|}\sum_{oj \in cj} dis[o_i][o_j]\right) \qquad (11)$$

Then, the most two closest clusters are merged until we reach the desired number of clusters.

Algorithm 4 : Merging close clusters

Inputs: set of initial clusters set C={$c_0, c_1, \dots, c_{n0-1}$}, the number of the correct clusters
Outputs: set of final clusters Cf

1: compute distance between all initial clusters using formula (11)
2: repeat
3: merge the two closest clusters
4: until number of clusters is reached

4 Experimentations and Clustering Validation

In order to evaluate our algorithm, experiments where conducted on 2.50GHz Intel ® Core ™ i5-3210M with 8Go RAM. All experiments were implemented in Java programming language.

We compared our algorithm with well-known clustering algorithms, K-means, DBSCAN (density-based spatial clustering of applications with noise), and SOM (Self organizing Map). The source code for those algorithms is used from Java Machine Learning Library JavaML [1]. As for, K-means and SOM, the obtained results are the mean of 10 runs.

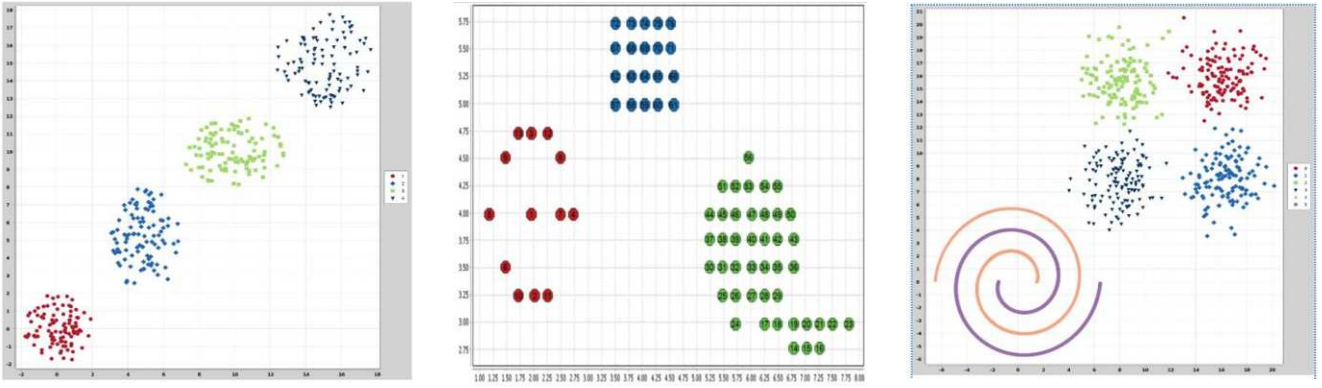
Experiments were conducted on five datasets, three of them are synthetic: Spharical_3_4[2], Dataset_3_2, and Spiralsquare[3]; the other two are real-world: Iris[4] and Wine[5] which are well known datasets for cluster evaluation:

(1) **Spharical_3_4:** contains 400 points on three dimensions distributed over four well separated spherical clusters.
(2) **Dataset_3_2:** contains 76 two-dimensional points distributes over three clusters of different shapes.
(3) **Spiralsquare:** contains 1500 two-dimensional points distributed over six clusters: four squared clusters slightly overlapping and two spiral clusters

[1] http://java-ml.sourceforge.net/
[2] Available at GitHub : https://github.com/deric/clustering-benchmark/tree/b47cdbb7028a61d632e2c63901f868e99444b350
[3] Available at GitHub : https://github.com/deric/clustering-benchmark/tree/b47cdbb7028a61d632e2c63901f868e99444b350
[4] Available at UCI Machine Learning Repository. : ftp://ftp.ics.uci.edu/pub/machine-learning-databases
[5] Available at UCI Machine Learning Repository. : ftp://ftp.ics.uci.edu/pub/machine-learning-databases

TABLE 1
Datasets description

Datasets		Instances	Attributes	Clusters
Synthetic datasets	**Spharical_3_4**	400	3	4 {100, 100, 100, 100}
	Dataset_3_2	76	2	3 {13, 43, 20}
	Spiralsquare	1500	2	6 {116, 134, 125, 125, 500, 500}
Real-world datasets	**Iris**	150	4	3 {50, 50, 50}
	Wine	178	13	3 {59, 71, 48}



(A) Spherical (B) Dataset_3_2 (C) Spiralsquare

FIG. 1. SYNTHETIC DATASETS, (A)SPHARICAL_3_4 (B) DATASET_3_2 (C) SPIRALSQUARE

(4) **Iris:** consists of three clusters of 50 four-dimensional points each; representing three categories of iris flowers (Setosa, Versicolor and Virginica). Setosa is a well separated cluster, while Versicolor and Virginica overlap.
(5) **Wine:** consists of 178 data points having 13 features each. This dataset is divided into three clusters representing three types of Wine.

Synthetic datasets are illustrated in Fig. 1. Characteristics of all datasets used in the experiments are shown in TABLE 1.

Among many different cluster evaluation metrics, and in the aim of validating our approach, we will be using:

- **Purity:** the purity is the percentage of the objects that were classified correctly, the purity of a cluster is calculated as follows:

$$P_j = \frac{1}{n_j} \text{MAX}_i(n_j^i) \quad (12)$$

Where n_j is the size of the cluster j, and n_j^i is the number of correctly assigned objects.
The overall purity the clustering is given by:

$$P = \sum_{j=1}^{k} \frac{n_j}{n} P_j \quad (13)$$

- Rand Index (RI): [21]The Rand index measures the percentage of decisions that are correct. It is calculated as follows:

$$RI = \frac{TP+TN}{TP+FP+FN+TN} \quad (14)$$

Where a true positive (TP) decision assigns two similar objects to the same cluster, a true negative (TN) decision assigns two dissimilar objects to different clusters. A (FP) decision assigns two dissimilar objects to the same cluster. A (FN) decision assigns two similar objects to different clusters.

- **Adjusted Rand Index:**[22] is the corrected-for-chance version of the Rand index, it is given as follows:

$$ARI = \frac{\sum_{lk}\binom{n_{lk}}{2} - \left[\sum_l \binom{n_l}{2} * \sum_k \binom{n_k}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_l \binom{n_l}{2} * \sum_k \binom{n_k}{2}\right] - \left[\sum_l \binom{n_l}{2} * \sum_k \binom{n_k}{2}\right] / \binom{n}{2}} \quad (15)$$

- **F-measure:** [23] F-Measure provides a single score that balances both the concerns of precision and recall in one number, it is given by the formula:

$$F_w = \frac{(W^2+1)*P*R}{W^2*P+R}, \quad (16)$$

Where : w is a positive real value, P is the precision, and R is the recall,

$$P = \frac{TP}{TP+FP} \quad (17)$$

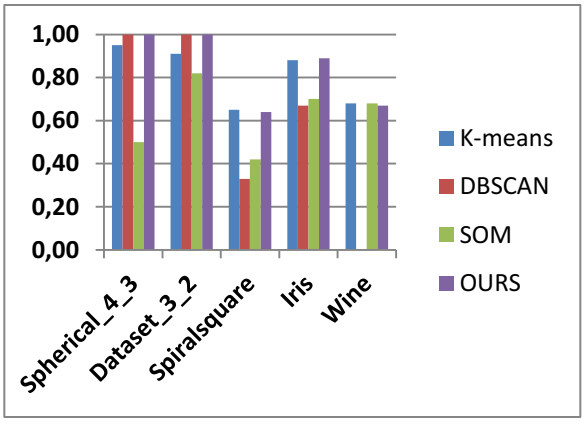


FIG. 2. PURITY VALUES RESULTED FROM VARIOUS ALGORITHMS ON DIFFERENT DATASETS. HIGH VALUE OF PURITY INDICATES BETTER CLUSTERING RESULT.

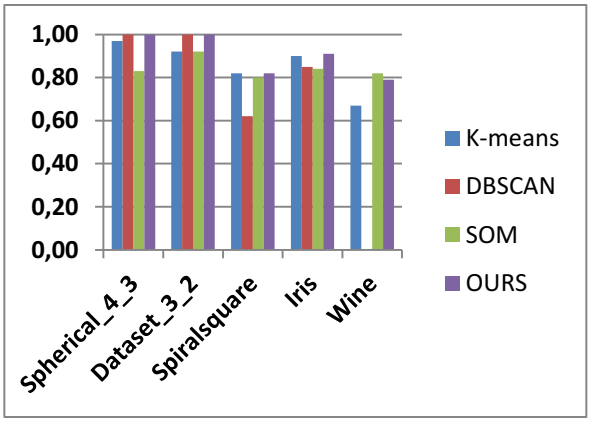


FIG. 3. RAND INDEX RESULTED FROM RUNNING VARIOUS ALGORITHMS ON DIFFERENT DATASETS. HIGH VALUE OF RAND INDEX INDICATES BETTER CLUSTERING RESULT.

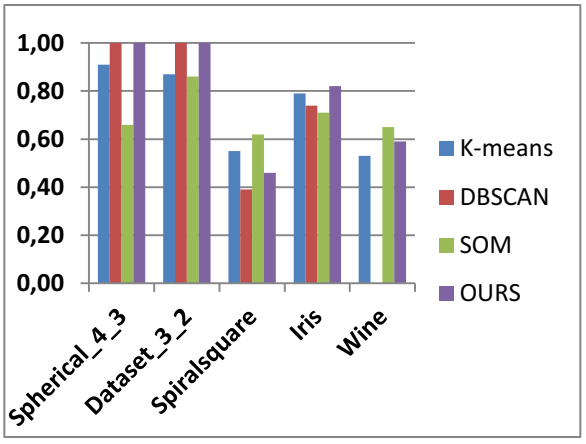


FIG. 4. F-MEASURE RESULTED FROM RUNNING VARIOUS ALGORITHMS ON DIFFERENT DATASETS. HIGH VALUE OF F-MEASURE INDICATES BETTER CLUSTERING RESULT.

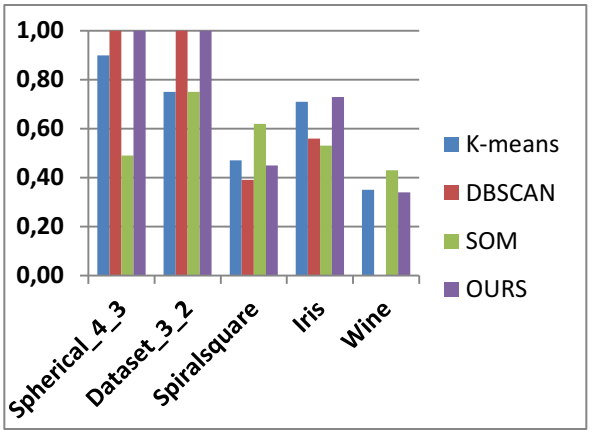


FIG. 5. ARI RESULTED FROM RUNNING VARIOUS ALGORITHMS ON DIFFERENT DATASETS. HIGH VALUE OF ARI INDICATES BETTER CLUSTERING RESULT.

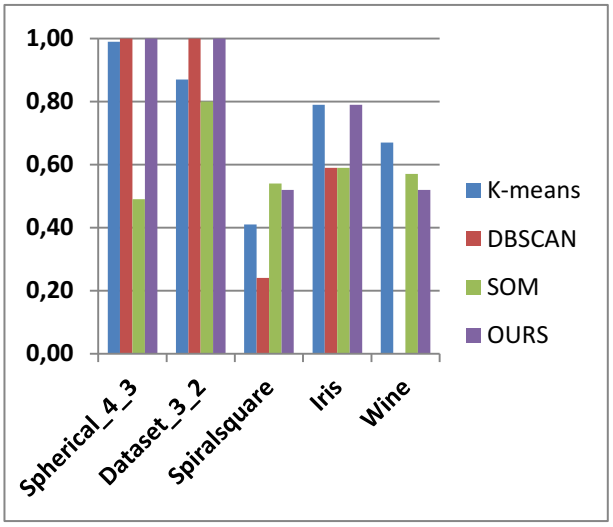


FIG. 6. PRECISION RESULTED FROM RUNNING VARIOUS ALGORITHMS ON DIFFERENT DATASETS. HIGH VALUE OF PRECISION INDICATES BETTER CLUSTERING RESULT.

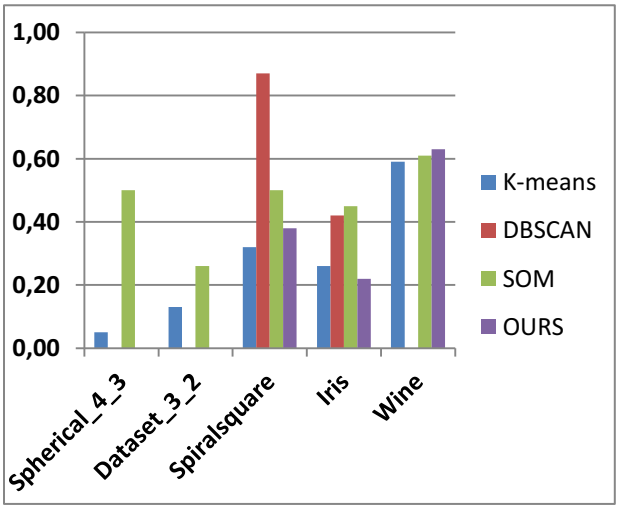


FIG. 7. ENTROPY RESULTED FROM RUNNING VARIOUS ALGORITHMS ON DIFFERENT DATASETS. LOW VALUE OF ENTROPY INDICATES BETTER CLUSTERING RESULT.

$$R = \frac{TP}{TP+FN} \quad (18)$$

- **Entropy:** [24] the entropy shows how the various classes of objects are distributed within each cluster, it is given by the following formula:

$$Entropy = \sum_{j=1}^{K} \frac{n_j}{n} E_j \quad (19)$$

Where E_j is the entropy of cluster j, it is given as follows:

$$E_j = -\frac{1}{\log k} \sum_{i=1}^{k} \frac{n_j^i}{n_j} \log \frac{n_j^i}{n_j} \quad (20)$$

Where: k is the number of clusters in the dataset.

The parameter L is the number of neighbors taken into consideration when calculating the connectivity. In our experiments setting we used the value 9 when the size of the dataset is less than 150; the value 14 when the size of the dataset is between 150 and 500; the value 28 when the dataset is bigger than 500.

4.1 Results and discussion

Clustering results obtained indicates that our algorithm MOCA-SM obtains overall good results for all datasets (Fig. 2, Fig. 3, FIG. 4, Fig. 5, Fig. 6, Fig. 7), noticing that other algorithms give good results for some datasets and less good results in others. The obtained clustering results are presented by dataset as follows:

(1) Spharical_3_4: for this dataset, our algorithm besides DBSCAN gives perfect clustering, as it is shown in the result figures. While K-means gets noticeably less good results, SOM gets the least good results for this dataset in all evaluation metrics.

(2) Dataset_3_2: similarly to the previous dataset, our algorithm and DBSCAN performed a perfect clustering, while K-means and SOM produced less good quality clustering; and it is valid for all evaluation metrics.

(3) **Spiralsquare:** for this challenging synthetic dataset, all four algorithms performed worse than the first two datasets. Although k-means slightly beats our algorithm, both of them obtained the best clustering results in terms of cluster purity, Rand Index, and entropy while DBSCAN and SOM get less than average results for this dataset. SOM gives the best precision result, closely followed by our algorithm. Regarding to F-measure and ARI, SOM gave best results.

(4) Iris: for this real-world dataset, our algorithm gives the best results in all evaluation metrics, closely followed by k-means, followed by DBSCAN and SOM which get close results for Iris dataset.

(5) Wine: for this dataset, SOM gets the best results closely followed by our algorithm for all metrics except for the precision where k-means takes the lead. for this dataset, DBSCAN results were not interpreted because the algorithm used eliminated 80% of the dataset points considering them as noise.

Comparison of the four algorithms for each clustering evaluation metric is presented in Fig. 2, Fig. 3, FIG. 4, Fig. 5, Fig. 6, Fig. 7.

As showed, our algorithm obtains generally good results for all tested datasets in contrast to the other well-known algorithms.

5 CONCLUSION AND FUTURE WORK

In this work, we have used congestion games with player-specific functions to model and to resolve the clustering problem. Using game theory tools allow a solid mathematical background to the proposed solution, which is considered to be one advantage of our work besides the good clustering results shown for different datasets.

The proposed algorithm operates on three phases where players are identified to play over a set of objects, or data points, and the aim of each plyer is to improve his own gains in terms of connectivity and R-square. In each game played, each player (or cluster-head) plays Nash equilibrium. When all players stop playing, the merging phase starts where cluster-heads decides to merge their clusters until having the final clusters.

Much further work is needed, especially in the 3rd phase to give our algorithm the ability of automatically deciding the number of clusters.

Although scalability results have not been treated in this paper, tests made until now made us think that the proposed clustering approach is very promising is this matter. Those ideas will surely be covered in the future work in order to enhance our approach and make it one of the most efficient clustering approaches.

ACKNOWLEDGMENT

The authors wish to thank the Insight Center for Data Analytics, UCD, Dublin, for inviting the first author as a visiting researcher while she was working on the development of this clustering approach.

REFERENCES

[1] E. Biernat and M. Lutz, *Data Science:fondamentaux et études des cas*. 2014.

[2] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Diane Cerr. Elsevier Inc., 2006.

[3] I. Gilboa and E. Zemel, "Nash and correlated equilibria: Some complexity considerations," *Games Econ. Behav.*, vol. 1, no. 1, pp. 80–93, 1989.

[4] T. Roughgarden, "Computing equilibria: A computational complexity perspective," *Econ. Theory*, vol. 42, no. 1, pp. 193–236, 2009.

[5] V. Conitzer and T. Sandholm, "New complexity results about Nash equilibria - tech report," *Games Econ. Behav.*, vol. 63, no. 2, pp. 621–641, 2008.

[6] C. Daskalakis, "The Complexity of Nash Equilibria," *Proc. 36th Ann. ACM Symp. Theory Comput.*, pp. 1–201, 2008.

[7] D. Xu and Y. Tian, "A Comprehensive Survey of Clustering Algorithms," *Ann. Data Sci.*, vol. 2, no. 2, pp. 165–193, 2015.

[8] C. C. Aggarwal and C. K. Reddy, *DATA Custering Algorithms and Applications*. 2013.

[9] A. Muthoo, M. J. Osborne, and A. Rubinstein, *A Course in Game Theory.*, vol. 63, no. 249. 1996.

[10] M. Wooldridge, "Does Game Theory Work ?," 2012.

[11] K. Leyton-Brown and Y. Shoham, "Essentials of Game Theory: A Concise Multidisciplinary Introduction," *Synth. Lect. Artif. Intell. Mach. Learn.*, vol. 2, no. 1, pp. 1–88, 2008.

[12] D. . Parkes and S. Seuken, "Game Theory I : Simultaneous-Move Games," 2013, pp. 17–41.

[13] J. Nash, "NON-COOPERATIVE GAMES," *Ann. Math. Second Ser. Ann. Math.*, vol. 54, no. 2, pp. 286–295, 1951.
[14] D. Easley and J. Kleinberg, "Games," in *Networks, crowds, and markets: Reasoning About a Highly Connected World*, Cambridge University Press, 2010, pp. 155–200.
[15] A. Fabrikant, C. Papadimitriou, and K. Talwar, "The complexity of pure Nash equilibria," in *STOC'04*, 2004, pp. 604–612.
[16] N. Helmi and G. Veisi, "A Multi-Modal Coevolutionary Algorithm for Finding All Nash Equilibria of a Multi-Player Normal Form Game," *Ubiquitous Inf. Technol. Appl.*, vol. 214, pp. 21–29, 2013.
[17] I. Milchtaich, "Congestion Games with Player-Specific Payoff Functions," *Games Econ. Behav.*, vol. 13, no. 1, pp. 111–124, 1996.
[18] R. W. Rosenthal, "A class of games possessing pure-strategy Nash equilibria," *Int. J. Game Theory*, vol. 2, no. 1, pp. 65–67, 1973.
[19] L. Gourv, "Congestion_games_-_Rosenthal.pdf," 2015.
[20] I. Heloulou, M. S. Radjef, and M. T. Kechadi, "Automatic multi-objective clustering based on game theory," *Expert Syst. Appl.*, vol. 67, pp. 32–48, 2017.
[21] W. Rand, "Objective criteria for the evaluation of clustering methods," *J. Am. Stat. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.
[22] L. Hubert, "Comparing partitions," *J. Classif.*, vol. 2, pp. 193–198, 1980.
[23] C. Rijsbergen, "Information retrieval," *London: Butter Worths*, 1979.
[24] Y. Zhao and G. Karypis, "Empirical and theoretical comparisons of selected criterion functions for document clustering," *Mach. Learn.*, vol. 55, pp. 311–331, 2004.

Dalila Kessira received the Engineer's degree in Industrial Computing from the University of Skikda, Algeria, in 2010; and she was awarded Magister degree (first level of the Post-graduation degree) in computer science from the University of Bejaia, Algeria, in 2015, where she worked on Digital Forensics readiness. She is currently working towards the Ph.D. degree at the Laboratory of Medical Informatics (LIMED), Department of Computer Science, University of Bejaia, Algeria. Her research work is focuses on is Data Mining techniques, Clustering and Game Theory applications. She is an assistant Lecturer in the same university from 2015 until now.

Professor Mohand Tahar Kechadi was awarded PhD and Master's degree - in Computer Science from University of Lille 1, France. He joined the UCD School of Computer Science (CS) in 1999. He is currently Professor of Computer Science at CS, UCD. His research interests span the areas of Data Mining, distributed data mining heterogeneous distributed systems, Grid and Cloud Computing, and digital forensics and cyber-crime investigations. Prof Kechadi has published over 260 research articles in refereed journals and conferences. He serves on the scientific committees for a number of international conferences and he organized and hosted one of the leading conferences in his area. He is currently an editorial board member of the Journal of Future Generation of Computer Systems and of IST Transactions of Applied Mathematics-Modelling and Simulation. He is a member of the communication of the ACM journal and IEEE computer society. He is regularly invited as a keynote speaker in international conferences or to give a seminar series in some Universities worldwide. The core and central focus of his research for the last decade is Data Mining techniques and models and their execution environments and applications (Forensic data, Medical data, distributed, cloud/Grid type services, etc.), Medical Data Mining: from knowledge to infrastructure, Digital Forensics and cyber-crime investigations, and Data-intensive (or driven) real-world applications.

From Entrepreneurial Leadership to Open Innovation: The Knowledge Management Role

Samah Chemli Horchani[#1], Mahmoud Zouaoui[#2]

[#] Management Department, Tunis El-Manar University, Faculty of Economics and Management Sciences of Tunis FSEGT, Laboratory of Innovation Strategy Entrepreneurship Finance and Economics LISEFE, Campus *Universitaire* Farhat Hached , B.P. 248 - *El Manar* II , 2092 *Tunis.*

***Abstract*— Entrepreneurial leadership is universally recognized and there are divergences in its effectiveness which suggests several promising areas of inquiry. The success of this leadership style depends on interrelations between leaders, followers and the Knowledge context. The article offers an interpretative reading of entrepreneurship and leadership theories to describe an innovative approach. It outlines the knowledge management importance. This work would be an opportunity for practitioners allowing them to discover the mechanisms and processes ensuring the maintenance of entrepreneurial intensity in an innovative company but also, opening the door to action when this intensity presents deficiencies.**

***Keywords*— Entrepreneurial leadership, open innovation, Knowledge management.**

I. INTRODUCTION

At this time of the twenty-first century, the economy description shows an uncertain environment characterized by the complexity of technologies and the arrival of generations Y and Z (Sahni, 2018). These generations are different from the old ones. They are more optimistic and confident. They take the risk and treat failure as an opportunity to learn (Tapcott, 1997; Bolton et al. 2013). These generations tend towards a new entrepreneurial management (Hoque et al., 2018). They are endowed with innovative ideas and want to create their own businesses. New entrepreneurs should no longer rely on natural and physical resources but on knowledge and sharing (Nonaka and Takeuchi, 1995; Alavi and Lidner, 2001; Wong and Aspinwall, 2005). Modern leadership styles must replace traditional leadership that is obsolete and incomplete (Luc and Le Saget, 2013). Indeed, an innovative idea does not mean that a company will progress and develop. Leadership is also necessary since it is the innovation engine (Donate and Sanchez de Pablo, 2015). However, the wealth of contributions on innovation, as well as on entrepreneurship and leadership, appears to be incomplete. The entrepreneurial leadership (EL) concept developed in this paper is a preliminary step which attempts to initiate more research in these directions and constitutes a new contribution in the knowledge management field within companies. The study examines how entrepreneurial leadership affects the innovation in a Knowledge context. It is aligned with new research on leadership and entrepreneurship combining orientation and action with characteristics and behaviors (Cogliser and Bigham, 2004; Renko et al., 2013). It outlines how innovation becomes a blend of knowledge and conditions, both internal and external. To address the research gap concerning conceptual development, we first provide a framework for entrepreneurial leadership. We develop a new conceptual model explaining in part how entrepreneurial leadership works within companies. The research offers advice to future entrepreneurs to promote their leadership at every level in their organizations. We first review existing research on entrepreneurial leadership and present key elements to discuss related constructs like entrepreneurial orientation and other leadership styles. After describing the field of entrepreneurial leadership, we specify three levels of studies that we associate with knowledge management and therefore try to detect their effects on open innovation.

II. LITERATURE REVIEW

A. Entrepreneurial leadership

The entrepreneurial leadership concept is a combination to explore both leadership and entrepreneurial behavior into a new leadership form (Gupta et al., 2004; Tarabishy et al., 2005). The

research path on entrepreneurship has coincided with research in economics and management. In the "managerial" entrepreneurial logic, the entrepreneurial genesis comes down to the resources exploitation thanks to the action of the "enterprising" entrepreneur in order to obtain a rent (Marchesnay, 2002). The new entrepreneurship approach supports the need for a resources arrangement which is a cognitive, an individual but alsoa socio-organizational construct. Then, it follows the commitment of different actors and the development of dynamic capacities around aptitudes, flexibility and creativity (Marchesnay, 2002). The company is thus embedded in its environment by building networks and by immersion in multiple networks (Marchesnay, 2002). However, the obstruction stems from the entrepreneurship definition, the entrepreneurial situations dissimilarity and the approaches diversity. During the research, several entrepreneurship definitions were presented. According to Coster (2009), entrepreneurship is the phenomenon of new opportunities emergence and exploitation to create economic or social value, driven and made possible by the man or the entrepreneur initiative and innovation dynamic , in interaction with his environment. For Lischeron (1991) entrepreneurship consists of having a role in changing and including new values in order to achieve objectives and create new opportunities. New knowledge is gaining importance, such as empowerment, maintaining organizational privacy and developing human resource systems. The entrepreneur is a leader capable of setting a vision and attracting others to achieve it. In the Marchesnay (1996) approach by the characteristics, the entrepreneur creates, detects and exploits opportunities. For these reasons, he must acquire skills enabling him to gather the resources necessary in order to detect the opportunities sources and in order to assess and exploit them (Janssen and Surlemont, 2009). In the same context, Leibenstein (1978) insinuates to the entrepreneur the role of making available the factors favoring the efficiency of "current production methods improvement or allowing the introduction of recent methods. This is made possible by the information flow in the market. For Hayek (1989) the knowledge acquisition and communication is the origin of adjustment to the market. The entrepreneur is then distinguished by vigilance with regard to the market imbalance (Kirzner, 1960 Cited by Ricketts and Kirzner, 1992); he must make a decisional arbitration to a competitive prices adjustment or by modifying the purchases and the sales planned according to the new prices on the market. Therefore, confidence in one's own judgment would be very important in ensuring a high level of profit (Knight, 1971). It should be noted that in the approach by the entrepreneur characteristics, everything revolves around the hero's characteristics who is the entrepreneur, that is to say, a typical profile search for a business creator entrepreneur. This approach is at the origin of the configuration analysis, based on the entrepreneurial firms' behavior, aiming to create companies groups classified according to their organizational forms (Short et al., 2008). Indeed, the configurations depend on the companies' entrepreneurial orientation that is to say on their innovativeness, propensities to take risks and their pro-activities (Randerson et al., 2011). Moreover, in configuration analysis, the entrepreneur's psychological traits influence certain enterprise's characteristics, such as entrepreneurial intent (Basso, 2006). Thus, an individual led by the need for accomplishment, having strong control over his conduct and destiny, and possessing a good capacity to succeed in tasks, will be more proactive, more predisposed to take risk and more likely to exploit entrepreneurial opportunities, than an individual with a low desire for accomplishment with little control over his behavior and a feeble ability to succeed in tasks (Randerson et al., 2011).

Some researchers have tried to present the entrepreneurial orientation through the leadership style exercised within the company (Chung Von, 2008). However, leadership is still an enigmatic theme (Jean-Michel Plane, 2015). This is due to the changes experienced by leadership due to globalization, digital transformation and the technology dynamics. This concept lacks a standard and formal definition (Conger, 1992) nevertheless a consensus remains present on its determining value and its main played role for the company's excellence. Thus, leadership is the investigation for an individual, labeled a leader, influencing a second individual, or group of individuals, labeled followers (Yukl and Van Fleet, 1992; Shao, 2018) to accomplish shared objectives (Yukl, 2010; Northouse, 2010). The leading entrepreneurs will put their proactivity into practice through new combinations of organizational capabilities to promote the reconfigurations and transactions necessary for their businesses development (Gupta

et al. 2004). The entrepreneurial leader must do the framing. Thanks to empowerment, he will give power by ensuring the balance between the desire to improve and his understanding of the individuals capacities. The leading entrepreneur combines ambitious goals with an insightful limits understanding (Brazeal and Herbert, 1999). By taking risks, the leading entrepreneur must absorb the uncertainty. This is done by absorbing the followers paralyzing effects and by building their confidence. The leader entrepreneur negotiates internal and external environmental relationships in order to clarify the path to follow and reduce the ability to change (Thompson, 1983). Information resources are crucial to carrying out this task (Daily and Dalton, 1993). An entrepreneurial leader is able to reshape the perception that individuals have of their own abilities by eliminating limitation self-imposed ideas (Gupta et al., 2004). The entrepreneurial leader empowerment and sharing , mold the team to deploy extraordinary energy and effort (Bandura, 1970). In the literature, the entrepreneurial leadership concept is seen through three essential levels (Hartog et al. 1999; Gupta et al., 2004): individual, organizational and social levels. The individual level relates to the leaders specifics and the roles they assume within the organization. The organizational level is about interactions within the organization. The social level relates to the exchange between the organization and its external environment.

B. Open innovation

C. Knowledge management

III. IMPACT OF ENTREPRENEURIAL LEADERSHIP ON OPEN INNOVATION

IV. RESEARCH METHODOLOGY

V. RESULTS

VI. CONCLUSION

REFERENCES