Cairo University

Institute of Statistical Studies and Research

The 48th Annual Conference on Statistics , Computer Sciences and Operation Research

Computer Sciences

23-26 Dec ,2013

Index

Computer Sciences

1	Semantically Enhancing Information Retrieval Results on Digital Libraries **Hassan. M.H , Abdellaltif.M.M , Salama.S.E and El Mahy.M.K**	**1-15**
2	An Aspect-Oriented Approach for SaaS Application Customization **Ashraf A.Shahin , Areeg Samir and Abdelaziz Khamis**	**16-30**
3	The Proposed Polygon Fuzzy Numbers for Fuzzy Multi Criteria Decision Making **Samah Bekheet , Ammar Mohammed and Hesham A.Hefny**	**31-45**
4	An Experiment for Performance Evaluation of Peer-to-Peer Mobile Web Services **Zaki A.M**	**46- 55**
5	An Enhanced Term Weighting Algorithm: A Fuzzy Ontology Approach **Zeinab E.Attia , Ahmed M. Gadallah and Hesham A.Hefny**	**56-64**
6	Handling OCR-Degraded Arabic Text: A Comprehensive Survey **Mariam A. Muhammad and Tarek A.Elghazaly**	**65-82**
7	Access Controlled Framework for Web-based Geographic Information System **Mennatallah H. Ibrahim , Nermin Hamza , Mohamed Nour El-Deen and Hesham A.Hefny**	**83-91**
8	Digital Change Detection in Burullus Protected Area Based on Temporal Multispectral Imageries **H.Farouk and B.Abdellatif**	**92-111**
9	Complexity Reduction of Fuzzy Rule Based System: Data-Sensitive Fuzzy Set Approach **Baha Eldin A , Ahmed M.Gadallah and Hesham A. Hefny**	**112 -123**
10	A Survey on Semantic Personalization of E-Learning Systems **Mohammed A.Moussa , Hesham A. Hefny and Akram Salah**	**124-140**
11	Manipulating Roughness Uncertainty in Optimization Problems **Assem A.Alsawy and Hesham A. Hefny**	**141-152**

Semantically Enhancing Information Retrieval Results on Digital Libraries

Hassan .M.H[1, 2], Abdellaltif.M.M[2, 3], Salama .S.E[2], ELmahy.M.K[4]

Abstract:

The Semantic Web is the new generation of the World Wide Web that enables people to share content beyond the boundaries of applications and websites. It has inspired many developers to create innovative semantic technologies and applications that play a major role to retrieve meaningful information intelligently from digital library. This paper provides a survey on the usage of semantic web in digital libraries and **presents** an ontology-based information retrieval using semantic indexing for digital library. This approach will provide semantic indexing with extracted and inferred information using domain ontology to achieve high retrieval performance and scalability for digital library. The indexing mechanism will be conducted using a free and open source information retrieval software library. In general this approach will deal with three issues in semantic search namely usability, scalability and retrieval performance. The performance of the system is improved considerably using domain-specific information extraction, inference and rules. Scalability is achieved by adapting a semantic indexing approach. The system will be implemented using Semantic Web technologies and its performance will evaluated against traditional systems as well as the query expansion methods.

Key words: semantic web, digital libraries, ontology

1. INTRODUCTION

Libraries were always the prime source of information for both students and scholars. Recently this information has been published online with digital library systems that provide efficient information discovery solutions to adapt the fast development of new technologies [1] they also have to cater the current generations of researchers. The research on the Semantic Web contributes to the digital libraries domain by supporting interoperability with formal semantics improving interlinking of information and encouraging users to contribute and share knowledge.

1-Hassan .M.H now pursuing her Master Degree from Helwan University, faculty of computers and Information, IS department
2- Helwan University, faculty of computers and Information, IS department
3- Dean of faculty of business, Delta University
4-MIEE.CEng, PhD- First Undersecretary- Chairman of STCE

Digital library community is already well organized and experienced in the skills required to best use of the Semantic Web because 'quick results ' could be gained, leveraging a large amount of material which is already catalogued [2]. The barriers in this area are more likely to be cultural with a tradition of centralized library services and copyright protected publications, coupled with a suspicion of the more open and anarchic world of the Web and slowing down adoption. Never the less so there is considerable interest in exploring the opportunities offered by the Semantic Web technologies that can help to deal with semantic heterogeneities and can give users a uniform interface for searching over distributed digital assets by maintaining machine-interpretable metadata.

This paper provides a survey on the usage of semantic web in digital libraries and **presents** an ontology-based information retrieval using semantic indexing for digital library. This approach will provide semantic indexing with extracted and inferred information using domain ontology to achieve high retrieval performance and scalability for digital library. The indexing mechanism will be conducted using a free and open source information retrieval software library. In general this approach will deal with three issues in semantic search namely usability, scalability and retrieval performance.

The performance of the system is improved considerably using domain-specific information extraction, inference and rules. Scalability is achieved by adapting a semantic indexing approach.

2. SEMANTIC WEB

This section aims to present an overview of semantic web [3] defines the semantic web as "a mesh of information linked up in such a way as to be easily processable by machines "on a global scale You can think of it as an efficient way of representing data on the World Wide Web, or as a globally linked database. People works on the Semantic Web base their work on the famous "semantic web tower" as illustrated in Figure 1 [4] this form is firstly presented by Tim Perners Lee [4]:

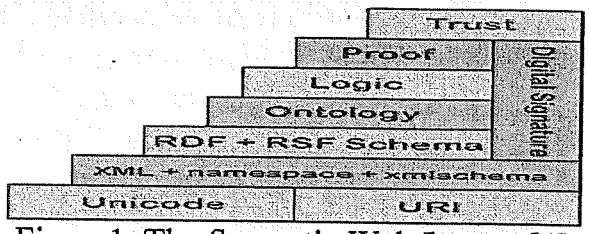


Figure1. The Semantic Web Layers [4]

A. Unicode and URI: Unicode the standard for computer character representation, XML: XML and its related standards, such as Namespaces, and Schemas

B. Resource Description Framework: RDF is the first layer of the Semantic Web proper. RDF is a simple metadata representation framework
C. Ontologies: Richer language for providing more complex constraints on the types of resources and their properties
D. Logic and Proof: An (automatic) reasoning system provided on top of the ontology structure to make new inferences
E. Trust: The final layer of the stack addresses issues of trust that the Semantic Web can support.

From a technical point of view the Semantic Web consists primarily of three technical standards which are RDF, SPARQL, and OWL [3]. The third technical standard OWL (Web Ontology Language) show that Ontology plays an important role in fulfilling semantic interoperability so ontology will discuss briefly in the next lines.

3. ONTOLOGY

It is a formal explicit description of concepts in a domain of discourse. [4]
Also its plays major rolls on semantic web as follow:

A. Helps to share common understanding of the structure of information among people and software agents
B. To enable reuse of domain knowledge to avoid "re-inventing the wheel" to introduce standards to allow interoperability
C. To make domain assumptions explicit
    - Easier to change domain assumptions (consider a genetics knowledge base).
    - Easier to understand and update legacy data.
D. To separate domain knowledge from the operational knowledge
E. Re-use domain and operational knowledge separately (e.g., Configuration based on constraints).

Also the process of developing ontology includes four steps as follow [5]:

A. Defining classes in the ontology.
B. Arranging the classes in a taxonomic (subclass–superclass) hierarchy.
C. Defining slots and describing allowed values for these slots, filling in the values for slots for instances
D. Then create a knowledge base by defining individual instances of these classes filling in specific slot value information and additional slot restrictions.

4. DIGITAL LIBRARY

A digital library is a library in which collections are stored in digital formats (CD, microform, or other media) [6] and accessible by computers the content may be stored locally or accessed remotely.

Needs new technology appears to store, retrieve and use of such a great and various amounts of informative texts, audio-visual resources, images that technology called digital library which provides:

- Expedite the systematic development of procedures to collect, store, and organize, information in digital form.
- Promote efficient delivery of information economically to all users.
- Encourage co-operative efforts in research resource, computing, and communication networks.
- Strengthen communication and collaboration between and among educational institutions.
- Take leadership role in the generation and dissemination of knowledge.

There are many limitations faces the digital library today that considers as barriers for maturity of digital libraries features. [7] These limitations motivate the researchers to provide great efforts in this field, some of these limitations are:

A. Coherence: Digital libraries should have a coherent, organizing principle, sorting materials by topic or by type of material.
B. Economic infrastructure - Although currently, many Internet sites are open, free of charge.
C. Searchability: The ability to find and use materials on the Internet is a major challenge for most users.
D. Reservation: By definition, libraries collect information for both current and future needs of users.
E. Retrieving large amounts of textual data quickly.
F. Allowing users to add annotations so that a reasoning capability exists.
G. Making text retrieval more specific.
H. Allow conclusions to be drawn by data on the Web and across Organizations.

Based on the previous limitations an extremely need for semantic digital library appeared to enhance the digital libraries performance and gives a good results fulfils users needs. Semantic web requirements can provide some functions for digital library to improve research processes result and these two main requirements divided into four approaches as follow:

A. new search paradigms for the information space
    - Ontology-based search approach.
    - Community-enabled browsing approach.
B. providing interoperability on the data level
    - Integrating metadata from various heterogeneous sources approach
    - Interconnecting different digital library systems approach.

5. SEMANTIC DIGITAL LIBRARIES REQUIREMENTS

Semantic technologies support more flexible information management than that offered by the classic digital libraries. Information about library resources can be composed from heterogeneous sources, including contributions from the communities of library users [8].

There are many literatures have been done to make Semantic Web technologies support and enhance digital libraries field and we divides this literature works according to the two main categories of building the Semantic Digital Libraries which consists of four approaches " as illustrated in (figure2) and the best ways to achieve them:

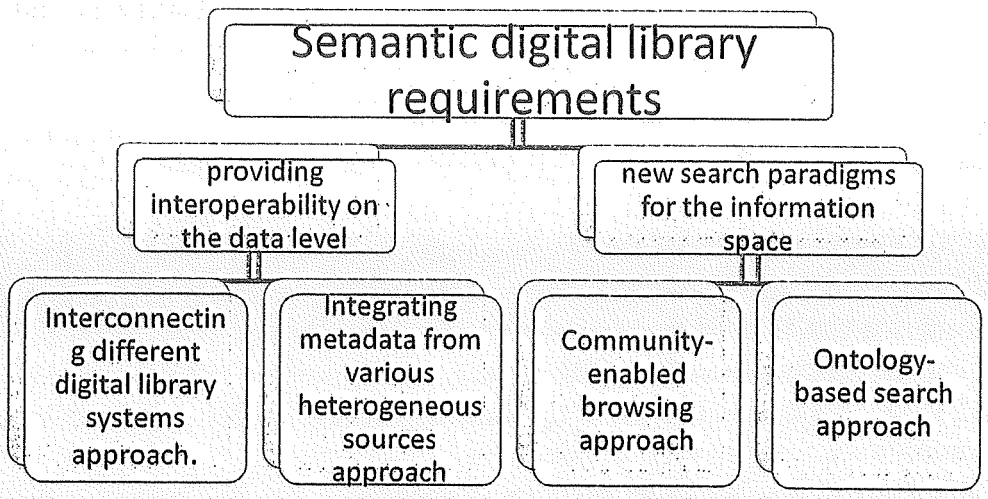


Figure2: categories *of the Semantic Digital Libraries*

A. New search paradigms for the information space

Search paradigms also called faceted search, it is a technique for accessing information organized according to a faceted classification system, allowing users to explore a collection of information by applying multiple filters [9]. The literatures will be discussed according to the two main approaches of search paradigms for the information space as follow:

- Ontology-based search approach

An ontology of the digital library evaluation domain named (DiLEO) was provided in [10] aims to reveal explicitly the main concepts of this domain and their correlations combines creatively and integrate several scientific paradigms, approaches, methods, techniques, and tools. This work demonstrates the added value features of the ontology which supports in comparative studies between different evaluation initiatives and the assistance in effective digital library evaluation planning.

The proposal in [11] Presents that digital libraries can perform the step from static information to dynamic knowledge spaces with the development of RDF, OWL, and SKOS by using ontologies in semantic

digital libraries. They propose to lift bibliographic metadata to a machine-interpretable semantic level by applying concepts defined in ontology then they applied ontologies to model the structural aspects of contents stored in digital library systems. Last they introduced ontologies as a mechanism to describe the knowledge of users and communities

The concept introduction on an analysis and of combining the areas of semantic web and web mining was presented in [12]. Emphasizing semantics in technologies, reasoning, content searching and social media

The research /13] explores the directions and the evolution of evaluation research for the period 2001–2011by discusses interesting patterns in the evaluation practices .The workshop in [14] focuses on the problems emerging from the marriage of "very-large" Digital Libraries and "very-large" Data Archives, in an attempt to manage, combine and interlink datasets with publications. Such problems include metadata definition, annotations, search, preservation, provenance, metadata aggregation, etc

A Model for Digital Libraries and a semantic query algorithm was provided in [9]. This model works in semantic ranking in Digital Libraries (DL Onto) Ontology-based Semantic Information Model for Digital Libraries by using an Ontology-based semantic query algorithm they a formal definition of the model and the functions on also gave a semantic query algorithm based on (DL Onto), then analyzed the semantic relativity of query results generated by this algorithm which can work well in semantic ranking in Digital Libraries.

- Community-enabled browsing approach

The proposal in [15] Presents SEKT approach which enables semantic queries to be specified, scheduled, and invoked periodically over the digital library's pre-indexed documents. Relevant knowledge is delivered proactively to users. Provide a knowledge sharing application enables a community of users to annotate web pages of interest, share those annotations, and explore the interests and topics of other users. The tool builds upon they ideas of user tagging and community links user tags leads to more formal ontology. A novel approach that provides personalized search by learning users' interests from historic behaviours and re-ranking search results by a Spreading Activation (SA) mode was provided in [16].

The research [17] explores issues that determine personalized services in academic libraries and also provides recommendations for successful lunching of such services in the library environment

The proposal in [18] introduces the notion of social search as a new concept, drawing upon the patterns of web search behaviour. It aims to: define social search; present taxonomy of social search; and propose a user-centred social search method. Various models that can be applied to online communities was described in [19] the proposal gives an overview of some of the new possibilities offered by such a layer in terms of data

querying and community analysis.The Living Document Project that aims to harness the collective knowledge within communities in digital libraries was presented in [20]. The Project makes it possible to enhance knowledge discovery and dissemination, facilitates interdisciplinary collaborations amongst readers also present a prototype that allows users to annotate content within digital libraries; the annotation schema is built upon the Annotation Ontology. Data is available as RDF making it possible to publish it as linked data and use SPARQL and SWRL for querying, reasoning, and processing. The LD (Living Document Project) approach offers an environment in which researchers harness the collective intelligence.

B. providing interoperability on the data level

Semantic interoperability provides interoperability at a highest level that provides the ability of two or more systems or elements to exchange information and to use the information that has been exchanged [21]. Semantic interoperability takes advantage of both the structuring of the data exchange and the codification of the data so that the receiving information technology systems can interpret the data to improve quality, safety, efficiency, and efficacy data delivery Integrating metadata from various heterogeneous sources which disuses here from its two main approaches as follow:

- Integrating metadata from various heterogeneous sources approach

Proposed A mediator system that provides the end-user with a single query interface to an integrated view of multiple heterogeneous data sources was provided in [22]. The system exploits the capabilities of the MOMIS integration system and the MILOS multimedia data management system. Each multimedia source is managed by an instance of MILOS

The research [23] on this work use different tools that have been developed in academic research and provide a conceptual model of how to implement these tools in relation to knowledge portals and warehouses.

A similarity measure using cross domain knowledge gathered from Wikipedia was provided in [24] they show that structure-based similarity measures are not suitable for chemical context searches and introduce a similarity measure combining entity- and context similarity

a metadata integration approach that provides the necessary features to fulfill three main requirements; was Presents in [25] support for any kind of data in and any type of interface to a data source, integration of heterogeneous metadata schemas, and location transparency.

A Metadata Management system that uses Semantic Web technology to handle the problem of heterogeneity on Web Service technology to achieve easy system integration presented in [26] their solution facilitates the development of integrated digital library systems, and lifting the burden of heterogeneous metadata management from application developers.

Ontology and integrated intelligent system architecture for search operation support system and its implementation platform have been developed in [27]. They also present ontology in Artificial Intelligent architecture for knowledge management in the Sevilla Digital Library system functions includes data processing and intelligent information retrieval this system combines RDF representation and CBR recommendation methodology to do code selection for the resources codes. integration of three structural value-adding methods which aim at reducing the semantic complexity represented in distributed DLs at several stages in the information retrieval process: query construction, search and ranking and re-ranking The integration of the model was done using Semantic Web technologies was proposed on [28] which should enhance further insights into the usage of those techniques.

-Interconnecting different digital library systems approach

a research methodology that can be used to map any human knowledge domain; proposed in[29] it is a novel and scientific method for producing comprehensive and systematic knowledge maps based on literary warrant a modern approach was presented in[30] where a modular environment provides an infrastructure of components to build DMLS applications upon. The Pundit which is a novel semantic annotation tool that allows users to create structured data while annotating Web pages relying on stand-off mark-up techniques propose in [31]. This Pundit provides support for different types of annotations, ranging from simple comments to semantic links to Web of data entities and fine granular cross-references and citations, Callimachus DL is presented in [32] a semantics-based digital library which provides faceted search, enhanced access possibilities and a proof-of-concept implementation. An interesting conceptual framework to introduce the ontology based knowledge management in the existing digital libraries presented in [33] which helps to transform any standard digital library into the Semantic Digital Library. A dedicated knowledge management unit is proposed here for the efficient and effective knowledge management within the digital library.

There are some drawbacks of this work as:

-No implemented prototype is introduced here since this work is still at conceptual and experimental level.

- Another major drawback of this framework is no security mechanism is introduced here but some researchers suggest the use of open source security tool to take the security measure which consider as a serious defect of their work. An approach combines a number of semantic technologies proposed in [34]; including ontology management, ontology learning and reasoning, keyword-type search and text classification they aim to allow the flexible and versatile answering of natural language questions on top of a digital library.

6. RESULTS OF THE SURVEY :

According to The observation made during the survey on literatures [10],[11],[12],[13],[14] that provide a techniques for accessing information organized according to a faceted classification system, they used Ontology-based search approaches that allowing users to explore a collection of information by applying multiple filters but they didn't addressed some aspects of semantic retrieval applying on decentralized heterogeneous and massive repository content of the digital that can be achieved by ranking algorithms to cope with large-scale information sources. Also In the same category [16], [17], [18] and [35] provide personalized and social search and services approaches from historic behaviours and re-ranking search results.These literatures provide a good search paradigms for the information space but they don't give their attention to **accessibility** and **usability** that is considered real gaps in semantic information retrieval from user perspective these gaps can bridge by combining the usability of keyword-based interfaces with the power of semantic technologies.

In general the study of these literatures shows that there are some open challenges such as **heterogeneity** that deal with unlimited number of domains. That can be achieved by generation of a SW gateway that collects and gives access to semantic metadata available online on digital library. Another open challenge is **scalability** in digital library environment that involves exploiting all the semantic metadata available online on digital library and to manage huge amounts of information in the form of unstructured content, these challenges can be overcome by Creation of scalable and flexible semantic indexing methods (Annotation).

For the rest of literature (Table1) give a comparison between them from some point of view such as (focus, semantic approach they use then define their tools and methods, finally the result of their work).

By / related work	Focus	Used approach	Semantic tools	Method	Result
G. Tsakonas, C. Papatheodorou	digital libraries evaluation domain named (DiLEO)	Ontology-based search approach	Ontology of digital library evaluation domain	Combines and integrate several scientific paradigms, approaches, methods, techniques, and tools	Demonstrates the added Devalue features of the ontology on digital libraries evaluation domain
K. Sebastian Ryszard, et.al.[8]	applying concepts defined in ontology	Ontology-based search approach	RDF, OWL, and SKOS	Applying concepts defined in ontology then applied ontologies to the structural of contents	Moving DI from static information to dynamic knowledge spaces
Z,Ming	(DLOnto) Semantic information Model	Ontology-based search approach	Ontology	Emphasizing semantics in technologies, reasoning, content searching and social media	semantic query algorithm
Y Sure, C Tempich, D Vrandecic, et.al	*SEKT* approach	Community-enabled browsing approach	Ontology	enables semantic queries to be specified, scheduled, and invoked periodically over the digital library's pre-indexed documents	community links user tags
L. Jael García Castro1, et.al	Living Document Project	Community-enabled browsing approach	RDF ,SPARQL and SWRL	earns the collective knowledge within communities in digital libraries by making it possible to enhance knowledge discovery and dissemination	Social tagging system
[16][17][18]	requisites of a personalization system	Community-enabled browsing approach	Ontology	Define social search; present taxonomy of social search; and propose a user-centred social search method.	Using ontologies promotes the integration of new services into existing
B Haslhofer, W Klas	Service Oriented Architectures (SOAs)	Integrating metadata from various heterogeneous sources approach	Ontology	Using cross domain knowledge gathered from Wikipedia and introduce a similarity measure combining entity- and context similarity	Querying the metadata sources and representing the results
H. Bernhard, H. Robert	Metadata Management system	Integrating metadata from various heterogeneous sources approach	Ontology	Handle the problem of heterogeneity on Web Service technology to achieve easy system integration	Facilitates the development of integrated digital library systems

M. Antonio, L. Carlos	an ontology in Artificial Intelligent architecture	Integrating metadata from various heterogeneous sources approach	RDF , Ontology	Provides the necessary features to fulfill three main requirements: support for any kind of data in and any type of interface to a data source, integration of heterogeneous metadata schemas, and location transparency	Introduce a prototype web based CBR retrieval system
P Mayr, P Mutschke, V Petras	integration of three structural value-adding methods	Integrating metadata from various heterogeneous sources approach	RDF , Ontology	Reducing the semantic complexity represented in distributed DLs at several stages in the information retrieval process: query construction, search and ranking and re-ranking	Sophisticated analytic tools and interfaces for the presentation of information
K. Sebastian Reszard , et.al	(Jerome DL) semantic digital library	Integrating metadata from various heterogeneous sources approach	RDF , Ontology	Build a mediator system that provides the end-user with a single query interface to an integrated view of multiple heterogeneous data sources	Improve scalability of the L2L network by limiting the required bandwidth in the nodes
S. Arup, M. Ujjal	ontology based knowledge management	Interconnecting different digital library systems approach	Ontology	Helps to transform any standard digital library into the Semantic Digital Library	No implemented prototype & no security mechanism
B. Stephan, et.al	Ontology based-structured natural language queries	Interconnecting different digital library systems approach	Ontology	Allow the flexible and versatile answering of natural language questions on top of a digital library.	Combination of different tools for natural language question interpretation, ontology learning, query answering & reasoning

Table 1 Comparison between the results of our survey on semantic digital library

Our approach builds on principles from [16] where a novel approach to provide personalized search by learning users' interests from historic behaviours and re-ranking search results. Users' interests are categorized by the level of regency to form user profiles. Spreading Activation model runs on the domain ontology incorporated with a newly defined relationship derived from the assumption of collaborative filtering.

In our approach we will provide additional features that can provide users with an easy to use query user interface in semantic-based technology, increase accessibility and usability of information retrieval from digital libraries, in addition increase information retrieval performance.

By combining semantic web technologies such as ontology development, information extraction, inference, semantic rules and semantic indexing and retrieval with the comfort of keyword-based search interface, the result will be one of the most user-friendly, high performance and scalable semantic retrieval system that can be applied on digital library so we will make an efforts towards increasing the retrieval performance while preserving the user-friendliness which will eventually come to the point of improving semantic searching with keyword-based interfaces.

As a way to achieve these features we **presents** an ontology-based information extraction and retrieval system and its application to digital library domain that deals with three issues in semantic search, namely, usability, scalability and retrieval performance.

On **a keyword-based** semantic retrieval approach the performance of the system is improved considerably using domain-specific information extraction, inference and rules Scalability will achieved by adapting a semantic indexing approach the system will be implemented using Semantic Web technologies and its performance also will be evaluated against traditional systems as well as the query expansion methods. Finally, show how to use semantic indexing to solve simple structural ambiguities.

7 .CONCLUSION AND FUTURE WORK:

This paper is a survey about using semantic web technologies with digital library, and provides a new approach to enhance information retrieval for digital library that using semantic indexing. Survey indicates that Semantic Web has been changed the way we used to design, develop and deploy technologies on digital libraries. The combination of XML, RDF, and ontologies makes it possible for computer programs to interpret digital objects based on their meaning and relationship to other concepts. Applying semantic web on digital libraries requires high degree of precision. Despite the huge number of researches on how semantic web affects digital libraries until now, there are still lack of more researches on usability of keyword-based search supported by semantic technologies to enhance and facilitate information retrieval from digital libraries that can

be achieved by using semantic indexing for efficient access to semantic knowledge.The big advantages that can be reached from using the Semantic Web is the easy access to the most Relevant and required Information, and the availability of contents with varied features and high quality. Although the fact that many researchers provides a great effort to make semantic web more helpful for digital library to cope with dramatic increase in information, and despite a number of institutions have undertaken the adoption of technology development for semantic web there are still several necessary problems need to be resolved and may Consider as a future point to start the work from as:

A. A major effort must be made to regulate the content of the digital libraries for storage and provide the necessary semantic web technology to find.
B. Must discover new methods differ from the view in the current digital libraries -based hypertext structure due to the high volume of information in them so the view of the content of the Semantic Web will be important to thinking about it because users will need easy retrieval to identify the information required.
C. Despite the intense enthusiasm for the Semantic Web and draws the whole world to use it but there is a serious problem, should pay attention to it namely (control and privacy), The advanced implementation of the Semantic Web will make it easy to see the information and impose censorship and guardianship
D. Current digital libraries facing the problem of multilingualism, it is necessary to find a radical solution to the problem by using the Semantic Web, Also must provide the tools and the means to gain access to information in many languages and independently of the original language of the party that offers the content as well as its users

References:

1- P. Goal," The Semantic Web: An Introduction and Issues International Journal of Engineering Research and Applications", (IJERA), 2012, pp.780-786
2-S. Sizov, "What Makes You Think That? The Semantic Web's Proof Layer", IEEE Intelligent Systems, November. IEEE Press, 2007,pp. ,22:94
3-A. Polleres. Semantic web technologies: From theory to standards. In 21st National Conference on Artificial Intelligence and Cognitive Science (AICS2010), Galway, Ireland, August 2010
4- F. Natalya, and L. Deborah " Ontology Development 101: A Guide To Creating Your First Ontology" ,Stanford University School of Medicine , (2001)
5-N. Madurai Meenachi , M. Sai Baba " A Survey on usage of Ontology in Different Domains ", International Journal of applied information system JAIS Journal 2012.
6-T.Mayank "Digital Libraries: Functionality, Usability, and Accessibility", Library Philosophy and Practice (e-journal), 2010.
7-C. Ángel García, et.al "Digital libraries and Web 3.0. The Callimachus DL approach" Computers in human behavior journal 2011, p.185.
8-K. Sebastian Ryszard "Digital Libraries of the Future Use of Semantic Web and Social Bookmarking to support E-learning in Digital Libraries ", Logic Group at Stanford University 2007
9-Z.Ming , D.Zhhong , D.Sukai and Y.Dongqing –" DLOnto: A Semantic Information Model for Digital Libraries" Semantics, Knowledge and Grid, 2005. SKG '05. First International Conference, 2005, page 16.
10-G Tsakonas, C Papatheodorou – " An ontological representation of the digital library evaluation domain" journal of the American Society for Information Science and Technology Volume 62, 2011, p.p. 1577–1593.
11-K. Sebastian Ryszard, K. Stefan, and Z. Lech " Jerome DL- Adding Semantic Web Technologies to Digital Libraries", Springer, Berlin Heidelberg ,Volume 3588, 2005, pp 716-725
12-P. Pablos, H Nigro, R Tennyson " Advancing Information Management Through Semantic Web Concepts and Ontologies
"- SEG Csaro - 2012 - dl.acm.org ISBN:1466624949
13-G Tsakonas, A Mitrelis "An exploration of the digital library evaluation literature based onan ontological representation
"- Journal of the American Society for Information Science and Technology,Volume 64, Issue 9, pages 1914–1926, September 2013
14-L Candela, P Manghi, Y Ioannidis *" Fourth workshop on very large digital libraries: on the marriage between very large digital libraries and very large data archives"- ACM SIGMOD Record, 2012*
15-Y Sure, C Tempich, D Vrandecic, et.al " SEKT Methodology: Initial Framework and Evaluation of Guidelines", Sekt deliverable d7.1.2, Institut AIFB, Universitaet Karlsruhe, 2006.
16-T Sun, M Zhang, F Yan, Z Deng " Personalized search in digital libraries via spreading activation model"- Web Intelligence and Agent Systems, 2013 - IOS Press
17- WM Beyene" Personalization and usage data in academic libraries: an exploratory study"- 2012 - oda.hio.no
18- M McDonnell, A Shiri " Social search: A taxonomy of, and a user-centred approach to, social web search"- (2011) ", Program: electronic library and information systems, Vol. 45 Iss: 1, pp.6 - 28
19-A Passant, S Kinsella, U Bojars" Understanding Online Communities by Using Semantic Web Technologies"- Hershey, PA : Information Science Reference, ©2011.

20- L.Jael García Castro, X .Olga. Giraldo. And C. Alexander Garcíathey-"Using the Annotation Ontology in Semantic Digital Libraries": Semantic Web Conference, ISWC, 2010
21-M.Grier,H.Robin. "Interoperability Solutions". *Interoperability.* Catalyst Communications. Retrieved 28 May 2011.
22- D Beneventano, C Gennaro, S Bergamaschi " A mediator-based approach for integrating heterogeneous multimedia sources"- January 2013, Volume 62, Issue 2, pp 427-450
23- H. Bernhard ,H. Robert "Metadata Management in a Heterogeneous Digital Library Web Engineering ", 4th International Conference, ICWE 2004
24- B Köhncke, WT Balke " Context-Sensitive Ranking Using Cross-Domain Knowledge for Chemical Digital Libraries- Research and Advanced Technology for Digital Libraries ,Lecture Notes in Computer Science Volume 8092, 2013, pp 285-296]
25- B Haslhofer, W Klas " A survey of techniques for achieving metadata interoperability", ACM Computing Surveys (CSUR)
26- *H. Bernhard, H. Robert*"Metadata Management in a HeterogeneousDigital Library" IOS Press 2005
27-M. Antonio, L. Carlospersonal" Intelligent Information Processing In A Digital Library Using Semantic Web", IADAT Journal 2011
28-P Mayr, P Mutschke, V Petras " Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking", Library Review, Vol. 57 Iss: 3, pp.213 – 224 , 2008
29- SH Nguyen, G Chowdhury Interpreting the knowledge map of digital library research - Journal of the American Society for Information Science and Technology, Volume 64, Issue 6, pages 1235–1258, June 2013
30- T Rosiek, W Sylwestrzak, A Nowiński" Infrastructural Approach to Modern Digital Library and Repository Management Systems"- Intelligent Tools for Building a Scientific Information Platform Studies in Computational Intelligence Volume 467, 2013, pp 111-128
31- M Grassi, C Morbidoni, M Nucci, S Fonda, and G Ledda " Pundit: Semantically Structured Annotations for Web Contents and Digital Libraries."- SDA, 2012
32- Á García-Crespo, JM Gómez" Digital libraries and web 3.0. The CallimachusDL approach- Computers in Human Behavior
Volume 27, Issue 4, July 2011, Pages 1424–1430
33-S.Arup, M.Ujjal "Ontology Based Knowledge Management Systems and Its Application to Semantic Digital Library Initiative", International Journal on Digital Libraries 2000, Volume 3, pp 237-248
34- B. Stephan, C. Philipp, D. Alistair, et.al "Ontology-based Question Answering for Digital Libraries", Research and Advanced Technology for Digital Libraries, 11th European Conference, ECDL 2007, Volume 4675, pp 14-25
35- V Uren, Y Lei, V Lopez, H Liu, E Motta, and M Giordanino" The Usability of Semantic search tools": The Knowledge Engineering journal Volume 22 Issue 4, December 2007 p 361-377

An Aspect-Oriented Approach for SaaS Application Customization

Ashraf A. Shahin, Areeg Samir, and Abdelaziz Khamis

Abstract

Multi-tenancy is one of the most important concepts for any Software as a Service (SaaS) application. Multi-tenant SaaS application serves a large number of tenants with one single application instance. Complex SaaS application that serves significant number of tenants could have a huge number of customizations with complicated relationships, which increases the customization complexity and reduces the customization understandability. Modeling such customizations, validating each tenant's customization, and adapting SaaS applications on the fly based on each tenant's requirements become very complex tasks. To mitigate these challenges, this research proposes an aspect-oriented approach that makes use of the Orthogonal Variability Model (OVM) and Metagraphs. The OVM is used to provide the tenants with simple and understandable customization model. A Metagraph-based algorithm has been developed to validate tenants' customizations. On the other hand, the aspect-oriented approach offers a high level of runtime adaptability. While the current validation algorithms partially deal with the variability and constraint dependencies, the proposed algorithm deals with all dependencies.

Keywords: Cloud computing, Software as a Service (SaaS), SaaS Application Customization, Aspect-Oriented Programming, Orthogonal Variability Model, Metagraph, AO4BPEL.

1. Introduction

Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly delivered with a minimal management effort or service provider interaction [1].

Software as a Service (SaaS) is a software delivery model in which software resources are accessed remotely by clients [2]. The SaaS delivery model is focused on bringing down the cost by offering the same instance of an application to as many customers, i.e. supporting multi-tenants. Multi-tenancy is one of the most important concepts for any SaaS application.

SaaS applications have to be customizable to fulfill the varying functional and quality requirements of individual tenants [3]. The elements of an application that need to be customized include: Graphical User Interface (GUI), workflow (business process logic), service selection and configuration, and data [4]. Several researches have attempted to support customization of these elements [2, 5, 6, 7, 8].

One of the most prominent approaches for achieving the tenant-specific customization is providing an application template with unspecified parts that can be customized by selecting components from a provided set of components [8, 9]. These unspecified parts are called customization points of an application [10].

Department of Computer and Information Sciences, Institute of Statistical Studies & Research, Cairo University, Egypt.

SaaS application customization research literature points out that there are four key concerns that need to be addressed: first, how to model the customization points and variations, second, how to describe the relationships among variations, third, how to validate customizations performed by tenants, and fourth, how to associate and disassociate variations to/from customization points during run-time [11].

Based on Orthogonal Variability Modeling (OVM) and Metagraphs, we propose an aspect-oriented approach for customizing SaaS applications which addresses the above concerns. OVM is used to model customization points and variations, and to describe the relationships among variations. A Metagraph-based algorithm has been developed to validate tenants' customizations. And finally, an aspect-oriented extension for the Business Process Execution Language (AO4BPEL) [12, 13] is used to associate and disassociate variations to/from customization points during run-time.

The remainder of the paper is organized as follows. In section 2, we give some background material on OVM, Metagraphs and AO4BPEL. Section 3 gives a short overview of related work. Section 4 presents our proposed approach to customize SaaS applications. In section 5, we evaluate the proposed approach. Finally, we conclude in section 6.

2. Background

2.1 Orthogonal Variability Modeling (OVM)

OVM is a proposal for documenting software product line variability [11]. In OVM only the variability of the product line is documented. In this model a *variation point* (*VP*) documents a variable item and a *variant* (*V*) documents the possible instances of a variable item. All VPs are related to at least one V and each V is related to at least one VP. Figure 1, taken from [11], shows for the graphical notation used in OVM.

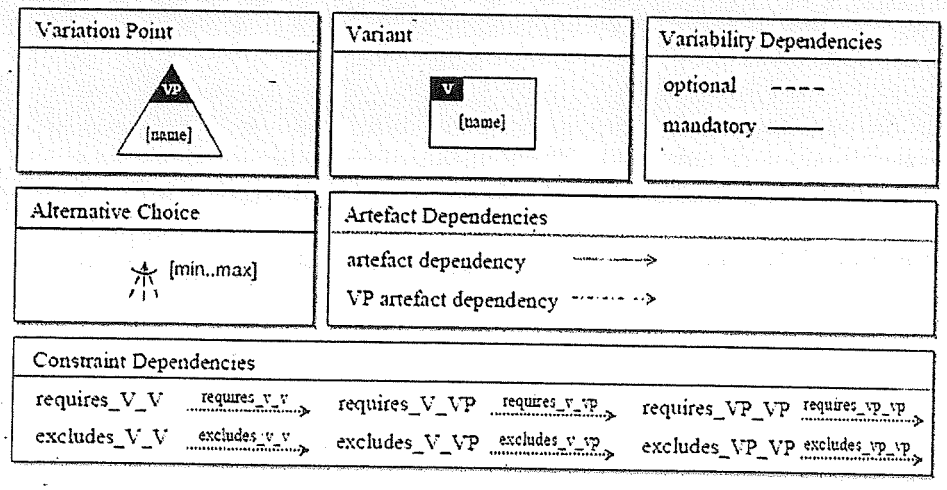


Figure 1. Graphical notation for OVM

In OVM, optional variants may be grouped in *alternative choices*. This group is associated to a cardinality [*min...max*]. Cardinality determines how many Vs may be chosen in an alternative choice, at least min and at most max Vs of the group.

In OVM, constraints between nodes are defined graphically. A constraint may be defined between Vs, VPs, and Vs and VPs. A constraint may be either *excludes* or *requires* constraint. An *excludes* constraint specifies a mutual exclusion, for instance, a variant *excludes* a VP means that if the variant is chosen to a specific product then the VP must not be bound, and vice versa. A *requires* constraint specifies an implication, for instance, a variant requires a VP means that always the variant is part of the product, and the VP must be also part of that product. Figure 2 depicts a simple example of an OVM inspired by the travel agent industry.

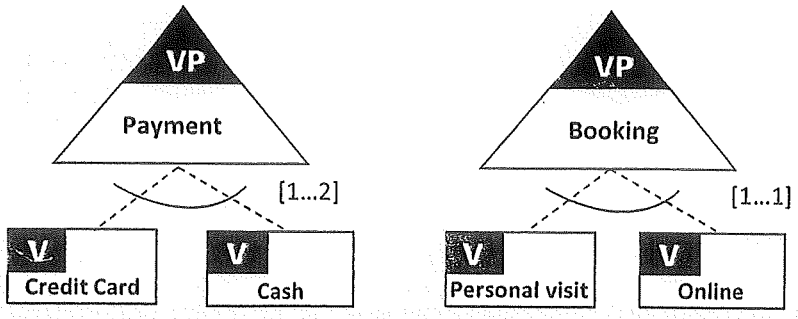


Figure 2. OVM Example: Travel agency industry

2.2 Metagraphs

A Metagraph is a graphical structure that represents directed relationships between sets of elements. The theory of Metagraphs is available in [14]. The remainder of this subsection includes a formal representation of a Metagraph, and an example of representing a simple business process as a Metagraph.

The *generating set* of a Metagraph is the set of elements X = {x1, x2, . . . , xn}, which represent variables of interest, and which occur in the edges of the Metagraph. An *edge* e in a Metagraph is a pair $e = \langle V_e, W_e \rangle \in E$ (where E is the set of edges) consisting of an *in-vertex* $V_e \subset X$ and an out-vertex $W_e \subset X$, each of which may contain any number of elements. A Metagraph $< X,E >$ is then a graphical construct specified by its generating set X and a set of edges E defined on the generating set.

An example of a Metagraph is shown in Figure 3. This Metagraph describes the activities and information elements in a travel agency business process. It contains three edges (labeled e_1, e_2, and e_3) that relate various information elements of concern in terms of specific activities. The edge e_1 represents the booking activity based upon the customer record; it results in either online booking or personal visit booking. The edge e_2 represents the payment activity based on the online booking; it results in credit card payment. The edge e_3 represents the payment activity based on the personal visit booking; it results in either credit card payment or cash payment.

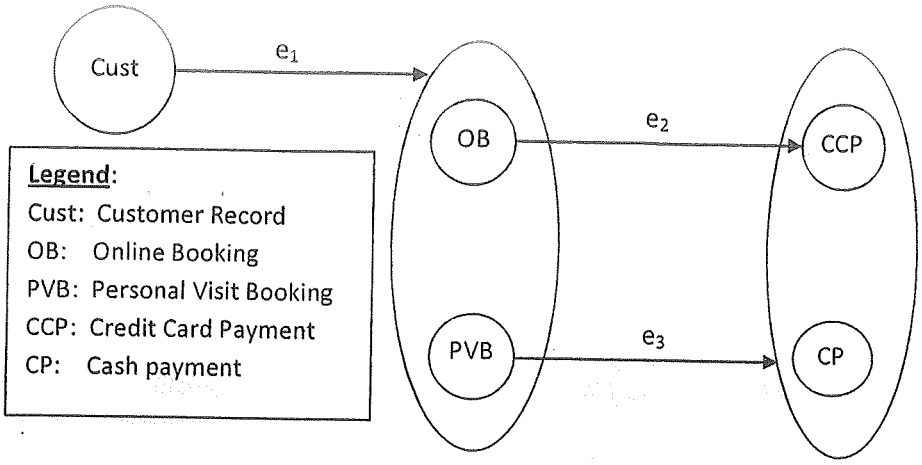


Figure 3. A Travel agency business process Metagraph

2.3 AO4BPEL

BPEL (Business Process Execution Language) is an XML-based language that enables task-sharing in a distributed computing environment. Using BPEL [12], a programmer formally describes a business process that will take place across the Web in such a way that any cooperating entity can perform one or more steps in the process the same way.

AO4BPEL [12, 13] is an aspect-oriented extension to BPEL, in which aspects can be (un)plugged into the composition process at runtime. Aspects enable the modularization of concerns that cut across multiple objects. Such concerns are often termed *crosscutting* concerns. An aspect uses *join points*, point *cuts* and *advices* in the modularization process. A join point is a point during the execution of a program. A Pointcut is a predicate that matches join points. An advice is an action taken by an aspect at a particular join point.

Figure 4, taken from [12], is an example of an aspect. The point cut of this aspect selects the flight search activities in the flight process and in the travel package process. The advice of this aspect contains an activity for starting a timer, which is executed before the join point as well as an activity for stopping a timer, which is executed after the join point.

3. Approach to SaaS Application Customization

SaaS applications are built following a service-oriented architecture (SOA) as it offers a flexible way for building new composite applications out of existing building blocks [3]. The layers of a SaaS application that need to be customized are: Graphical User Interface (GUI), business process logic (workflow), service, and data [4]. In our proposed approach to customize SaaS applications, we will be concerned only with the workflow and service layers.

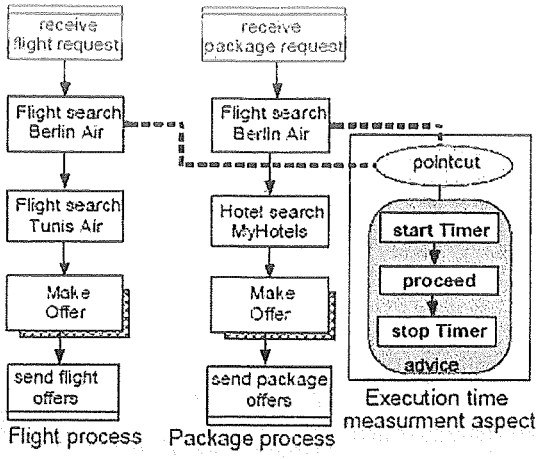


Figure 4. An Aspect Example: Execution time measurement

In our proposed approach, we use OVM to model customizations in workflow and service layers. A Metagraph-based algorithm has been developed to validate tenant-level customizations. AO4BPEL is used to associate and disassociate variations to/from customization points during run-time. Finally, we present a framework for our proposed approach. In the following subsections we demonstrate our approach through a simplified example from the Travel Agency Domain that includes only the booking and payment activities.

3.1 Modeling Workflow Customization

In SOA development, a workflow refers to how two or more web services interact [15]. Workflow design often follows the template design pattern. The application designers who design workflows need to define customizable places where the customization is possible. Each customizable place can be replaced by one or more sub-workflow templates. Each sub-workflow contains an atomic or a composite web service, and it can still have customizable places.

The application to model in this paper is taken from the travel agency domain. It contains two activities: booking and payment processes. Figure 5 shows a customizable workflow for this application. It contains two customizable places. Each customizable place in the workflow of our application can be replaced by two sub-workflows. These sub-workflows are given in Figure 6.

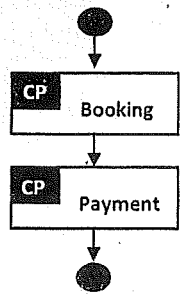


Figure 5. A customizable workflow

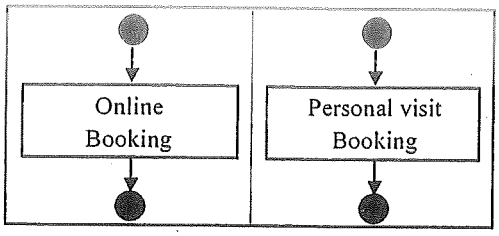


Figure 6.a Sub-workflows for booking

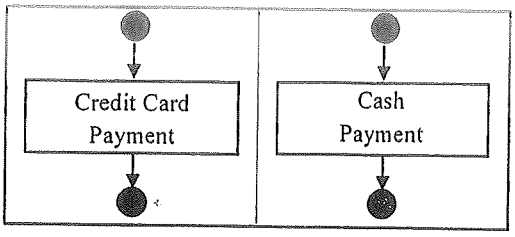


Figure 6.b Sub-workflows for payment

In multi-tenant applications, to provide tenants with understandable customization with constraint dependencies, we model the customizable workflow into OVM. Figure 7 shows the OVM of the customizable workflow for the simplified travel agency application. The customizable places in the workflow are modeled as variation points, and each sub-workflow is modeled as variant.

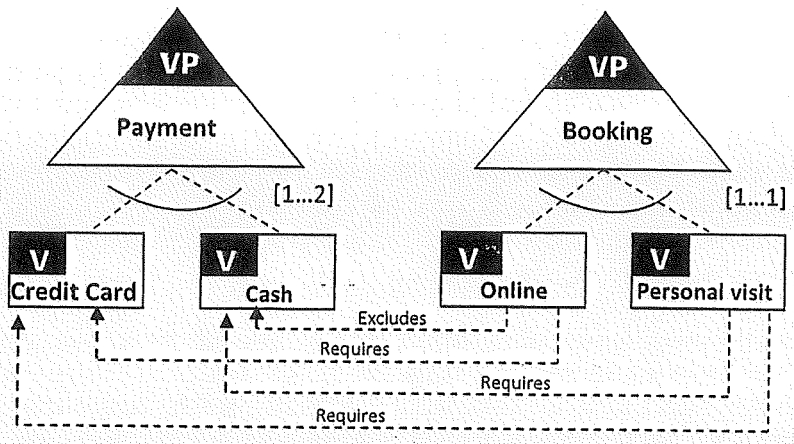


Figure 7. OVM of customizable workflow

3.2 Modeling Service Customization

A web service is the main building block to build a SaaS application. A web service customization is the ability to extend, configure, and change web service behavior to achieve particular requirements [16]. To enable SaaS application venders to build highly customizable SaaS applications, the involved web services need to be highly customizable.

To customize a web service, providers need to identify commonalities and variations across the scope of their SaaS application. Identified commonalities are realized as *core services* that exist in all customized applications. Identified variations are realized as *variant services*. Tenants customize web services by selecting one or more of these variation web services.

We model a customizable service into OVM. The customizable service is modeled as a variation point, and each of the core and variant services is modeled as a variant. Figure 8 models a customizable service with one mandatory core service and two optional variant services.

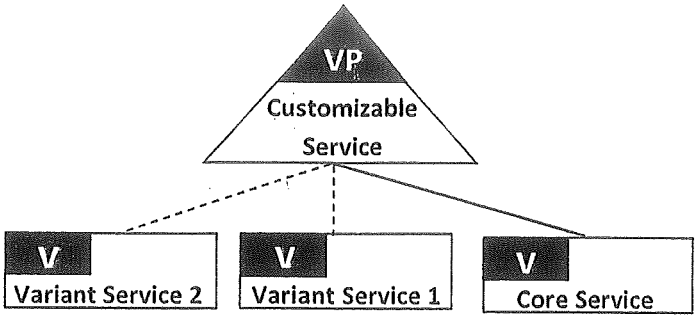


Figure 8: OVM of customizable service

3.3 Tenant-Level Customization Validation

In this section we develop a Metagraph algorithm to validate the tenant-level customization that has been modeled as OVM. We start with showing how to map OVMs to Metagraphs, and then we introduce our validation algorithm. Finally, we apply the algorithm on our case study.

3.3.1 Mapping OVMs to Metagraphs

All variables of interests in OVM (Variation Points and Variants) compose the generating set of the Metagraph that models OVM. Thus, all variation points and variants in OVM are mapped to *vertices* in the Metagraph. The variability and constraint dependencies in OVM are mapped to *edges*, labeled with qualitative attributes, defined on the generating set. Following this mapping procedure, the OVM shown in Figure 7 is mapped to the Metagraph shown in Figure 9.

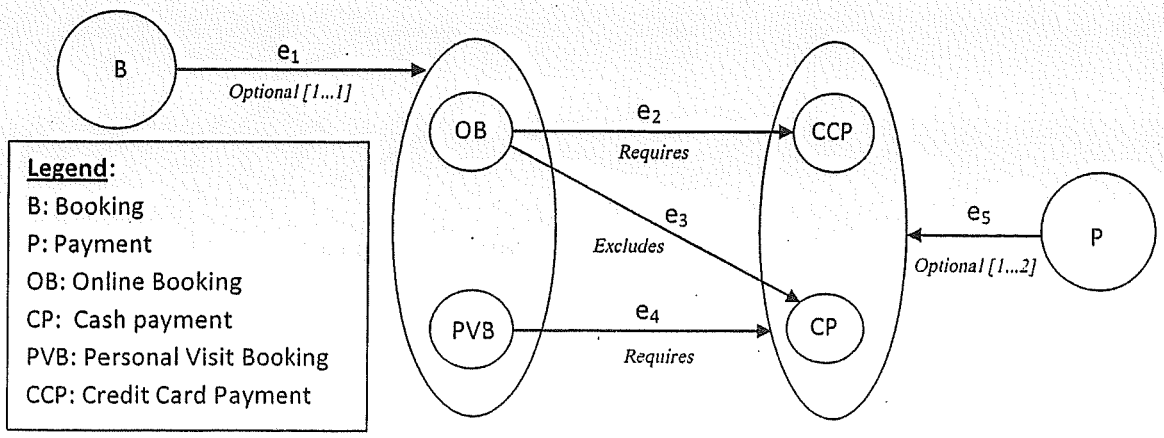


Figure 9. Metagraph representation of the OVM in figure 7

In order to use the above Metagraph in our validation algorithm, we need an algebraic representation of the attributes that are represented visually in the Metagraph. This done by adding each attribute as an additional element to the generating set, and then includes that

element in the in-vertex of the edges to which it applies. Applying this procedure to the Metagraph in Figure 9, we get an equivalent Metagraph, shown in Figure 10, in which the qualitative attributes are represented as vertices. The cardinality associated with the alternative choices will be represented as a matrix called the cardinality matrix and shown in Figure 11.

Finally, in order to manipulate the Metagraph, in Figure 10, by our validation algorithm we construct an adjacency matrix representation of the Metagraph. The adjacency matrix A of a Metagraph is a square matrix with one row and one column for each element in the generating set X. The ij^{th} element of A, denoted a_{ij} , is a set of triples, one for each edge *e* connecting x_i to x_j. Each triple is of the form $<CI_e, CO_e, e>$, in which CI_e is the co-input of x_i in *e* and CO_e is the co-output of x_j in e. Figure 12 shows the adjacency matrix of the Metagraph in Figure 10.

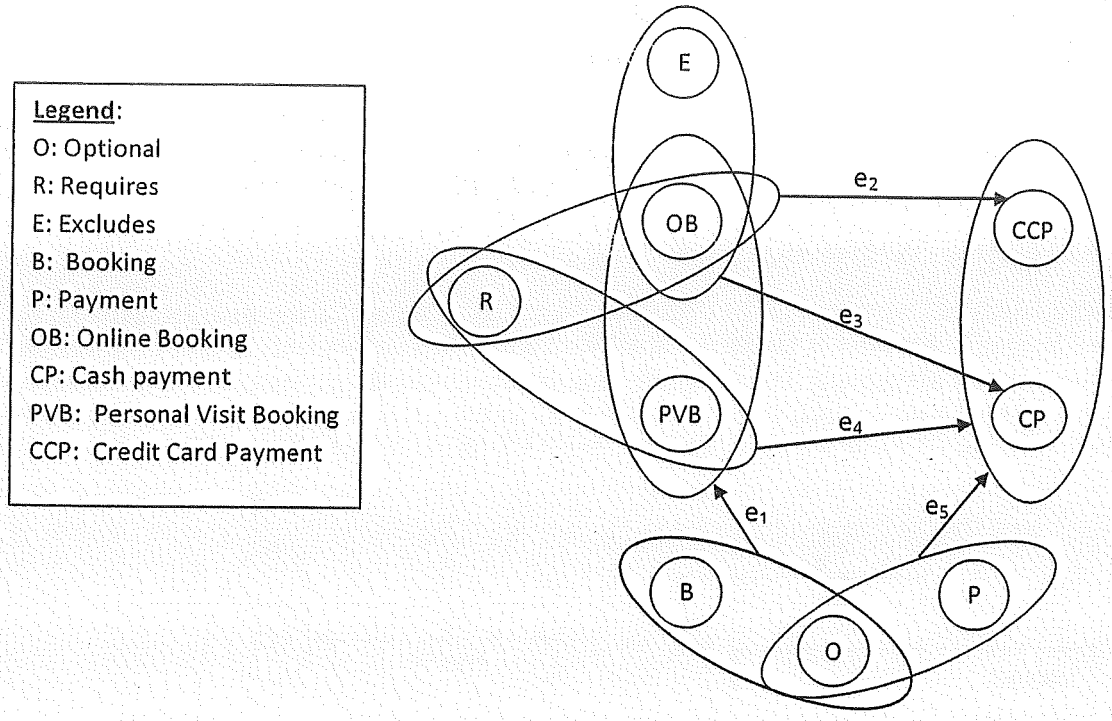


Figure 10. Vertex representation of qualitative attributes

	e_1	e_5
Min	1	1
Max	1	2

Figure 11. The cardinality matrix

	B	OB	PVB	CCP	CP	P	R	E	O
B	φ	<{O},{PVB},e_1>	<{O},{OB},e_1>	φ	φ	φ	φ	φ	φ
OB	φ	φ	φ	<{R}, φ,e_2>	<{E}, φ,e_3>	φ	φ	φ	φ
PVB	φ	φ	φ	φ	<{R}, φ,e_4>	φ	φ	φ	φ
CCP	φ	φ	φ	φ	φ	φ	φ	φ	φ
CP	φ	φ	φ	φ	φ	φ	φ	φ	φ
P	φ	φ	φ	<{O}, {CP},e_3>	<{O},{CCP},e_5>	φ	φ	φ	φ
R	φ	φ	φ	<{OB}, φ,e_2>	<{PVB}, φ,e_4>	φ	φ	φ	φ
E	φ	φ	φ	φ	<{OB}, φ,e_3>	φ	φ	φ	φ
O	φ	<{B},{PVB},e_1>	<{B},{OB},e_1>	φ	<{P}, {CCP},e_5>	φ	φ	φ	φ

Figure 12. The adjacency matrix of the Metagraph in Figure 10

3.3.2 Customization Validation Algorithm:

Tenant-level customization is performed by assigning variant(s) to each variation point. In our algorithm, the variation points in our application are referred to by CP (Customizable Points) and the variants are referred to by V.

The inputs to our validation algorithm are: (1) M = <X, E>: the Metagraph that models the customizable SaaS application, (2) ICP: Initial Customizable Points, (3) R: the cardinality matrix, and (4) C = {c_1, c_2, …, c_n}: the set of customizations performed by the tenant. Each c_i is represented as an order pair <cp_i, v_i>, where the cp_i represents a customizable point, and v_i represents the variant assigned to cp_i.

The outputs of our validation algorithm are: (1) M_T = <X_T, E_T>: the Metagraph that models the valid tenant-level customizations, (2) IC: an invalid customization made by the tenant, if found, (3) VF: the validation flag that indicates whether the customization performed by the tenant is valid or not, and (4) CF: the completeness flag that indicates whether the customization performed by the tenant is complete or not.

Figure 13 shows the steps of our validation algorithm. Lines 1-3 construct the adjacency matrix, A, of the Metagraph M, initialize the Metagraph M_T with ICP and all mandatory edges whose in-vertices belong to ICP, and construct the adjacency matrix, A_T, of the Metagraph M_T.

Lines 4-20 identify the invalid customizations made by the tenant, if any found. Lines 21-25 add the valid customization made by the tenant to the Metagraph M_T. Line 26 updates the adjacency matrix, A_T, of the Metagraph M_T.

A customization instance represented by the pair <cp_i, v_i> is valid if:

- All the customizable points involved in required customizations, with customization variant v_i, must be selected before selecting v_i.
- The customization variant v_i should not be excluded from the alternative choices for the customizable point cp_i.
- The customizable point cp_i should be selected before selecting the variant v_i.

- The customization variant v_i should be one of the customization variants that customize the customizable point cp_i.
- Adding the given customization instance should not exceeds the maximum cardinality associated with the alternative choices for the customizable point cp_i.

Lines 27-29 make sure that all mandatory customization variants have been associated to the relevant the customization points in X_T. Lines 30-32 make sure that the minimum cardinality associated to any alternative choice group is achieved. Finally, Line 33 returns two flags: the flag CF indicates that the customization made by the tenant is complete. The flag VF indicates that all customization instances made by the tenant are valid.

Begin
1: Construct the adjacency matrix, A, of the Metagraph M.
2: Initialize M_T: (X_T = ICP and E_T = All *mandatory* edges whose in-vertices belong to ICP)
3: Construct the adjacency matrix, A_T of the Metagraph M_T.
4: **for each** element c_i in the customization set C. (c_i = <cp_i, v_i>)
5: **for each** non-empty element a[v_i][j] in the row v_i of the matrix A
6: **for each** triple <CI_e, CO_e, e> in a[v_i][j]
7: **if** the *requires-vertex* $\in CI_e$ and $x_j \notin X_T$, where x_j refers to the vertex corresponding to the column j.
8: CF = incomplete, IC = c_i , VF = invalid, and **return**
9: **for each** non-empty element a[j][v_i] in the column v_i of the matrix A
10: **for each** triple <CI_e, CO_e, e> in a[j][v_i],
11: **if** the *excludes-vertex* $\in CI_e$ and $x_j \in X_T$, where x_j refers to the vertex corresponding to the row j.
12: CF = incomplete, IC = c_i , VF = invalid, and **return**
13: **if** $cp_i \notin X_T$
14 CF = incomplete, IC = c_i , VF = invalid, and **return**
15: **if** the element a[cp_i][v_i] in A is empty
16: CF = incomplete, IC = c_i , VF = invalid, and **return**
17: e′ = e, where e is the edge in the triple $\langle CI_e, CO_e, e\rangle \in$ a[cpi][vi] in A, such that the *optional-vertex* $\in CI_e$
18. m = The union of the co-outputs of all triples $\langle CI_e, CO_e, e\rangle$ in the row cp_i of A_T, such that e = e′
19: **if** m = R[max][e′]
20: CF = incomplete, IC = c_i , VF = invalid, and **return**
21: $X_T = X_T \cup \{v_i\}$
22: **If** $e' \in E_T$ and $e' = \langle V_{e'}, W_{e'}\rangle$
23: $W_{e'} = W_{e'} \cup \{vi\}$
24: **else**
25: $E_T = E_T \cup \{e'\}$, where $e' = \langle V_{e'}, \{vi\}\rangle$
26: Update A_T
27: **for each** triple $\langle CI_e, CO_e, e\rangle$ in the *mandatory* row in the matrix A
28: **if** $CI_e \subseteq X_T$ and $CO_e \not\subseteq X_T$
29: CF = incomplete, VF = invalid and **return**
30: **for each** triple $\langle CI_e, CO_e, e\rangle$ in the *optional* row in the matrix A_T
31: **if** $|CO_e| \leq$ R[min][e],
32: CF = incomplete, VF = invalid and **return**
33: CF = complete, VF = valid and **return**
End

Figure 13. Customization validation algorithm

3.4 Runtime Customization using AO4BPEL

Each *customization point* in a SaaS application has one or more *customization variants*. Tenants customize the SaaS application by selecting a customization variant and associate it to the customization point. Tenants may disassociate a customization variant from a customization point and replace it with another customization variant. In our approach, we use AO4BPEL to enable modular implementation of these customizations and dynamic adaptation during run-time.

By using AO4BPEL, customization variants can be composed to or decomposed from their customization points during run-time without stopping, rebinding, recompiling, or even restarting the applications. In our approach, every customization point is designed as a BPEL process, and each customization variant is designed as an aspect. All BPEL processes and the all aspects are developed by the *SaaS application developer*.

For the application presented in this paper, we need to design two BPEL processes for the booking and payment customization points. In addition, we need to design four aspects for the customization variants (online booking, personal booking, cash payment, and credit card payment). Based on the customization made by the tenant, the AO4BPEL engine dynamically weaves the selected aspects (customization variants) with their corresponding processes (customization points).

All the files related to the BPEL processes, the aspects, the weaving process, and the involved web services (source code, WSDLs, and file descriptors) are stored in three stores: *Process Store, Aspect Store, and Service Store*.

3.5 The Framework of the proposed Customization Approach

This subsection summarizes the activities involved in customizing a SaaS application. Figure 14 describes the customization scenarios based on our customization approach. The *Customization-Validation* unit, which is an implementation of our customization validation algorithm, takes as inputs: (1) the SaaS customization data prepared by the SaaS application developer and stored in the *SaaS-Customization-Data* database, (2) the tenants' customization set prepared by the tenants' administrators, and (3) the previous validated customizations made by the same tenant's administrator and stored in the *Validated-Customization-Data* database. The output of the validation process will be stored in the *Validated-Customization-Data* database.

The SaaS application developer implements all valid customizable points as processes and stores them the *Process Store* unit. The developer also implements all valid customization variants as aspects and stores them in the *Aspect Store* unit. All web services that are related to the developed processes and aspects are stored in the *Service Store* unit.

The *Validation UI* unit enables the tenants' administrators to define their customization sets, and send a request to the *Customization-Validation* unit to validate these sets. It also enables the tenants' administrators to make the necessary actions if the customization set is incomplete, or if any customization instance is invalid.

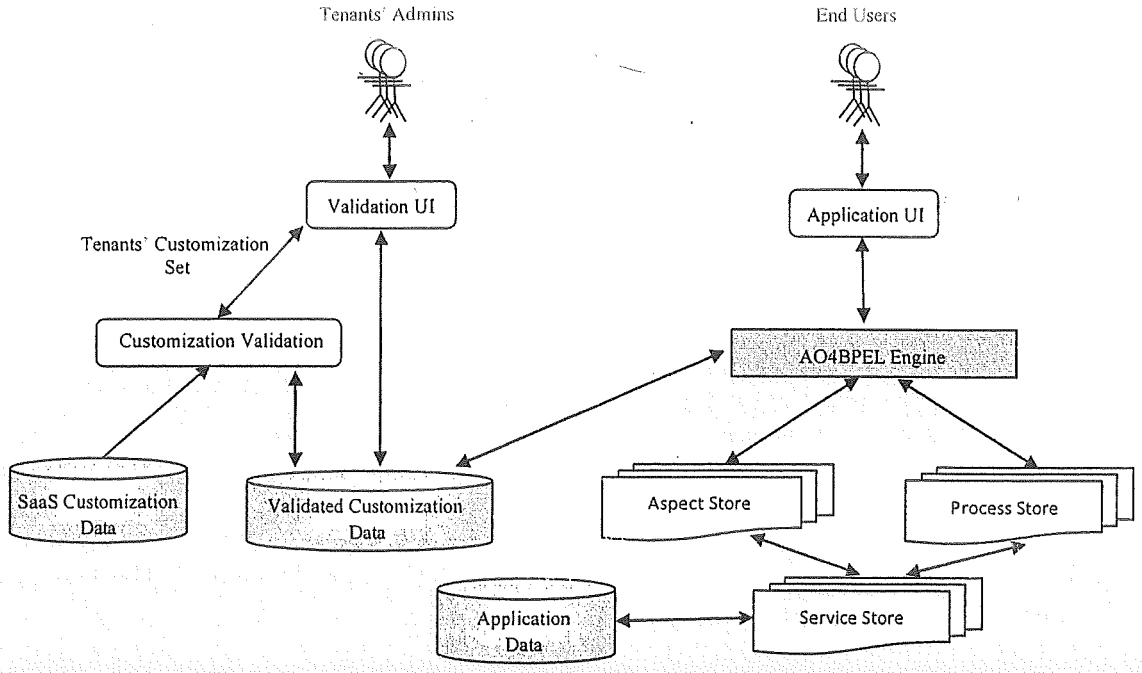


Figure 14. The Framework of our Proposed Approach

Based on a request from an end user, the AO4BPEL engine retrieves the relevant validation customization data for its tenant from the *Validated-Customization-Data* database, and then waves the corresponding process and aspects, and finally, the engine loads the waved process to perform the user request.

4. Comparison with Related Work

Customization could be performed in two different approaches: *Source-Code Based* or *Composition Based*. In the source-code based customization, SaaS applications are customized by adding and integrating new source code with the customizable application. Examples of research papers that follow this approach include [2, 5, 7, 17, 18].

The source-code based customization approach although it provides tenants with flexibility in the customization process, it suffers from some drawbacks including: (1) The tenant administrator should be aware of the implementation details of the SaaS application, (2) Allowing the tenant administrator to add and integrate source code may violate the security regulations of the application, and (3) The source code added by the tenant should be revised when the SaaS application developer upgrade the application.

In the composition based approach, SaaS applications are customized by composing variant components, selected from a provided set of components, to the customizable application. In addition to our proposed approach in this paper, examples of other research papers that follow this approach include [3, 8, 10, 15,19, 20].

As mentioned earlier in the introduction, there are four key concerns that need to be addressed in the SaaS application customization: first, how to model the customization points and variations, second, how to describe the relationships among variations, third, how to validate customizations performed by tenants, and fourth, how to associate and disassociate variations to/from customization points during run-time.

In our approach we dealt with all the above mentioned concerns by providing the tenant with simple and understandable customization model, developing a customization validation algorithm, and making use of the aspect-oriented approach to handle the runtime customization. On the other hand, many of the other related work partially addressed these concerns as discussed below.

In [10], the authors dealt with three concerns by using Metagraphs to model customization points, variants, and their relationships. They also proposed a validation algorithm to validate customizations made by tenants. However, the proposed model dealt only with the "*requires*" relationship. On the other hand, Metagraph-based modeling is very difficult for tenants to understand, especially in large applications with thousands of variants.

The authors in [4] handled only one concern that is modeling customization points, and variants using ontology-based intelligent customization framework. In [15], the authors also dealt with only one concern that is modeling customization points, and variants using OVM. They also guided the tenant through the customization process.

Finally, in [3], the authors dealt with three concerns by using OVM to model customization points, variants, and their relationships. To avoid unpredictable customizations, tenants are guided through the customization process. However, in large applications with thousands of variants, the guiding process they proposed will not be very helpful for the tenant.

5. Conclusion and Future Work

This paper discussed the concerns that need to be addressed to customize SaaS applications, and then proposed an aspect-oriented approach to address these concerns. OVM is used to model customizations in workflow and service layers. A Metagraph-based algorithm has been developed to validate tenant-level customizations. AO4BPEL is used to associate and disassociate variations to/from customization points during run-time. We demonstrated our approach through a simplified example from the Travel Agency Domain. Finally, we compare our proposed approach with others in related work.

We can conclude that the proposed architecture not only avoided the drawbacks of the source-code based customization but also addressed all the key concerns in the SaaS application customization, and all the variability and constraint dependencies. In the future, we will implement the framework of the proposed approach, and provide detailed analysis and evaluation of the framework.

References

[1] P. Mell and T. Grance, "The NIST definition of cloud computing," *National Institute of Standards and Technology*, p. 7, Jan. 2011.

[2] W. Lee and M. Choi, "A multi-tenant web application framework for SaaS," in *2012 IEEE 5th International Conference on Cloud Computing (CLOUD)*, 2012, pp. 970–971.

[3] R. Mietzner, A. Metzger, F. Leymann, and K. Pohl, "Variability modeling to support customization and deployment of multi-tenant-aware software as a service applications," in *2009. PESOS 2009. ICSE Workshop on Principles of Engineering Service Oriented Systems*, 2009, pp. 18–25.

[4] W. Tsai, Q. Shao, and W. Li, "OIC: Ontology-based intelligent customization framework for SaaS," in *2010 IEEE International Conference on Service-Oriented Computing and Applications (SOCA)*, 2010, pp. 1–8.

[5] P. Aghera, S. Chaudhary, and V. Kumar, "An approach to build multi-tenant SaaS application with monitoring and SLA," in *2012 International Conference on Communication Systems and Network Technologies (CSNT)*, 2012, pp. 658–661.

[6] M. Pathirage, S. Perera, I. Kumara, and S. Weerawarana, "A multi-tenant architecture for business process executions," in *2011 IEEE International Conference on Web Services (ICWS)*, 2011, pp. 121–128.

[7] J. Lee, S. Kang, and S. J. Hur, "Web-based development framework for customizing java-based business logic of SaaS application," in *2012 14th International Conference on Advanced Communication Technology (ICACT)*, 2012, pp. 1310–1313.

[8] H. Moens, E. Truyen, S. Walraven, W. Joosen, B. Dhoedt, and F. De Turck, "Developing and managing customizable software as a service using feature model conversion," in *2012 IEEE Network Operations and Management Symposium (NOMS)*, 2012, pp. 1295–1302.

[9] W. Sun, X. Zhang, C. J. Guo, P. Sun, and H. Su, "Software as a service: Configuration and customization perspectives," in *2008. SERVICES-2. IEEE, Congress on Services Part II*, 2008, pp. 18–25.

[10] C. Lizhen, W. Haiyang, J. Lin, and H. Pu, "Customization modeling based on metagraph for multi-tenant applications," in *2010 5th International Conference on Pervasive Computing and Applications (ICPCA)*, 2010, pp. 255–260.

[11] K. Pohl, G. Böckle, and F. J. v. d. Linden, *Software Product Line Engineering: Foundations, Principles and Techniques*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[12] A. Charfi, "Aspect-oriented workflow languages: AO4BPEL and applications," Ph.D. dissertation, Technische Universität Darmstadt, Germany, 2007.

[13] A. Look, "Expressive scoping and pointcut mechanisms for aspect-oriented web service composition," Master's thesis, Technische Universität Darmstadt, Germany, 2011.

[14] A. Basu and R. W. Blanning, *Metagraphs and Their Applications*. Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA, 2006.

[15] W.-T. Tsai and X. Sun, "SaaS multi-tenant application customization," in *2013 IEEE 7th International Symposium on Service Oriented System Engineering (SOSE)*, 2013, pp. 1–12.

[16] H. Jegadeesan, and S. Balasubramaniam, "A Method to Support Variability of Enterprise Services on the Cloud," in *2009 IEEE International Conference on Cloud Computing (CLOUD '09)*, 2009, pp. 117 – 124.

[17] X. Zhou, L. Yi, and Y. Liu, "A collaborative requirement elicitation technique for SaaS applications," in *2011 IEEE International Conference on Service Operations, Logistics, and Informatics (SOLI)*, 2011, pp. 83–88.

[18] L. Kong, Q. Li, and X. Zheng, "A novel model supporting customization sharing in SaaS applications," in *2010 International Conference on Multimedia Information Networking and Security (MINES)*, 2010, pp. 225–229.

[19] J. Park, M. Moon, and K. Yeom, "Variability modeling to develop flexible service-oriented applications," *Journal of Systems Science and Systems Engineering*, vol. 20, no. 2, pp. 193–216, 2011.

[20] Q. Li, S. Liu, and Y. Pan, "A cooperative construction approach for SaaS applications," in *2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2012, pp. 398–403.

The Proposed Polygon Fuzzy Numbers

For Fuzzy Multi Criteria Decision Making

Samah Bekheet[1], Ammar Mohammed[2], Hesham A. Hefny[3]

Abstract

The Fuzzy Decision making plays an important role in our life application especially under fuzzy environment which has rapidly changing in its circumstances. Fuzzy multi criteria decision making problems need a big effort to determine the weight of each criteria as expert's opinions and also may there is a need to determine the weight of each expert's opinion to determining the weight of each alternatives. Those are depending on linguistic values, which must be converted to numerical values to calculate the mathematical operations, which lead us to use the fuzzy numbers. Many of researchers used triangle fuzzy number system and some others used trapezoidal fuzzy number system. Each fuzzy number system has its own rules, which may be difficult to convert from system to others to check correctly and accuracy. This paper introduces a generalized polygon fuzzy system method, which can used any fuzzy number system as triangle, trapezoidal, pentagon, hexagon, heptagon, octagon and etc.

Keywords: Fuzzy multi criteria decision making, Polygon fuzzy numbers, Level sets Lemma, Centroied Theorem, fuzzy geometric mean.

1. Introduction

Decision making is the procedure to find the best alternatives among a set of feasible alternatives and also ranking them as their priorities. Under a fuzzy environment, fuzzy multi criteria decision making need fuzzy numbers system to express the expert's opinions since these opinions are linguistic values as excellent, very good, good, bad and so on.

[1] Institute of Statistical Studies and Research, Cairo University, Egypt samahbekheet@yahoo.com

[2] Dr. of Computer Science, Institute of Statistical Studies and Research, Cairo University, Egypt.ammar@eu.ed.eg

[3] Prof. of Computer Science and Head of Department of Computer and Information Sciences, Institute of Statistical Studies and Research, Cairo University, Egypt.hahefny@ieee.erg

These linguistic values need to any fuzzy numbers to evaluate their calculations as interval, triangle and trapezoidal fuzzy numbers as used before on most of fuzzy multi criteria decision making problems. But it is not easy to change from one system to another to be sure the results are very critical and there is no method to calculate the evaluations with different fuzzy numbers systems altogether to insure the results are congruent and very accurate. In this paper we will introduce a general polygon fuzzy number system which can converted from any polygon to others and also can solve the problem with different polygon fuzzy systems as triangle, trapezoidal, pentagon, hexagon, heptagon and octagon and so on to be sure these results are very critical, accurate and precise [1][2][3], This method has few equations and calculates to all different fuzzy systems by the same way depending on the number of vertices. We can consider this method as a general case of triangle and trapezoidal or any other polygon fuzzy numbers, which introduce a wonderful technique to check correctly and accuracy. The rest of the paper is organized as follows: Section II defines the problem, Section III introduces the proposed solution, Section IV illustrates definitions and classification of information, Section V introduced a numerical example, Section VI presents the conclusion, finally Section VII lists the references.

2. Problem definition

Solving fuzzy multi criteria decision making problems used triangle fuzzy numbers at most or trapezoidal fuzzy numbers rarely which depending on ranking the criteria or alternative as the highest weight, The weight of each criteria or alternative can be calculated by many different ways. One of the most important ways is evaluating the centre of area (centroied) of the triangle fuzzy shape or trapezoidal fuzzy shape [1][9]. But by introducing more polygons fuzzy shapes as pentagon, hexagon, heptagon and octagon. There is a problem at calculating the centre area of each shape simply and accurately, and also there is no method to solve a problem with different fuzzy numbers systems.

3. Proposed solution

Related to the importance of polygons fuzzy shapes which introduce the generalized fuzzy numbers, since it plays a big role in comparing the results and in measuring their correctness and accuracy, this paper introduces an improved theorem to calculate the centre of area of any polygon with n vertices called the centroied theorem and also introduces a standardized method to solve the decision making problem with different fuzzy numbers systems.

3.1. Level set lemma

Under identifying universe of discourse [0,1] and defining different polygon shapes with "V" vertices Figure1, it must to determine the number of level sets of polygon "N" defined in (1),which also is the number of different heights with lengths 1/n,2/n ,....,1

$$N= (V-1)/2 \tag{1}$$

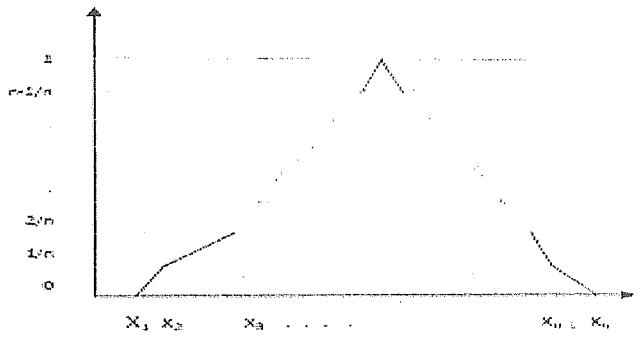


Figure 1 Level set characteristic

3.2. Centroied Theorem

The (Centroied) weight of polygon area with defined in (2), which distributed in small polygon, where the weight of area of each one defined in (3) and the area of small polygon defined in (4).

$$\bar{x} = \frac{\sum_{i=1}^{n} \overline{X}_i A_i}{\sum_{i=1}^{n} x_i} \tag{2}$$

$$\overline{X}_i = \frac{1}{3}[(x_i + x_{i-1}) + (\frac{x_i y_i + x_{i-1} y_{i-1}}{y_i + y_{i-1}})] \tag{3}$$

$$A_i = \frac{(x_i - x_{i-1})(y_i + y_{i-1})}{2} \tag{4}$$

3.3 Theorem Proof

By distributed the area of the polygon with n vertices into n-1 of small polygons as described in Figure 2 to calculate the centroied of each Ai as illustrated in Figure 3.

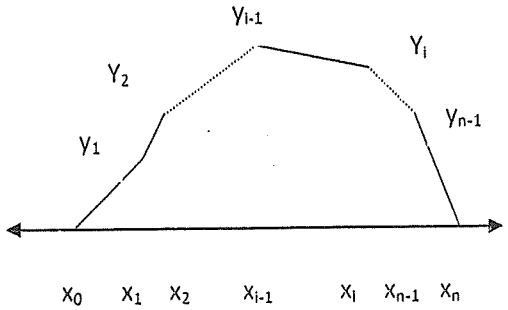


Figure 2 Polygon characterized

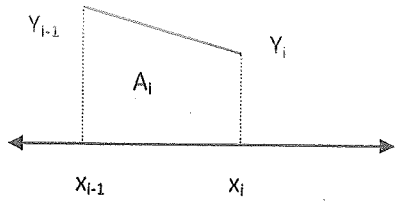


Figure 3 Small polygon area

$$\bar{x}_i = \frac{\int_{x_{i-1}}^{x_i} x\, f_i(x)\, dx}{\int_{x_{i-1}}^{x_i} f_i(x)\, dx} = \frac{I_1}{I_2}$$

$$\frac{f_i(x) - y_{i-1}}{x - x_{i-1}} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}$$

$$f_i(x) = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}(x - x_{i-1}) + y_{i-1}$$

$$= \frac{y_i - y_{i-1}}{x_i - x_{i-1}} x - \frac{y_i - y_{i-1}}{x_i - x_{i-1}} x_{i-1} + y_{i-1}$$

$$= \frac{\Delta y_i}{\Delta x_i} x + y_{i-1} - \frac{\Delta y_i}{\Delta x_i} x_{i-1}$$

$$f_i(x) = \alpha_i x + \beta_i$$

where $\alpha_i = \frac{\Delta y_i}{\Delta x_i}$, $\beta_i = y_{i-1} - \frac{\Delta y_i}{\Delta x_i} x_{i-1}$

Then $I_1 = \int_{x_{i-1}}^{x_i} x(\alpha_i x + \beta_i)\, dx$

$$= \frac{\alpha_i x^3}{3}\Big|_{x_{i-1}}^{x_i} + \frac{\beta_i x^2}{2}\Big|_{x_{i-1}}^{x_i}$$

$$= \frac{\alpha_i}{3}[x_i^3 - x_{i-1}^3] + \frac{\beta_i}{2}[x_i^2 - x_{i-1}^2]$$

$$= \frac{(x_i - x_{i-1})}{6}[(x_i + x_{i-1})(y_i + y_{i-1}) + x_i y_i + x_{i-1} y_{i-1}]$$

Also $I_2 = \int_{x_{i-1}}^{x_i} (\alpha_i x + \beta_i)\, dx$

$$= \frac{(x_i - x_{i-1})}{2}(y_i + y_{i-1})$$

So $$\bar{x}_i = \frac{\frac{(x_i - x_{i-1})}{6}[(x_i + x_{i-1})(y_i + y_{i-1}) + x_i y_i + x_{i-1} y_{i-1}]}{\frac{(x_i - x_{i-1})}{2}(y_i + y_{i-1})}$$

$$\bar{x}_i = \frac{1}{3}[(x_i + x_{i-1}) + (\frac{x_i y_i + x_{i-1} y_{i-1}}{y_i + y_{i-1}})$$

4 .DEFINITIONS AND CLASSIFICATION OF INFORMATION

By applying the previous improved theorem, we can define each polygon fuzzy shape as illustrated.

4.1. Triangle fuzzy numbers

Under universe of discourse [0,1] as described in Figure 4 has relationship $\mu_t(x)$ defined in (5) [8][9] where $x_0 \le x_1 \le x_2$, $y_0 = y_2 = 0 and\ y_1 = 1$.

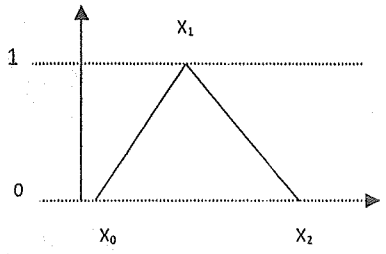


Figure 4 Triangle characterized.

$$\mu_t(x) = \begin{cases} \dfrac{x - x_0}{x_1 - x_0} & x_0 \le x \le x_1 \\ \dfrac{x_2 - x}{x_2 - x_1} & x_1 \le x \le x_2 \end{cases} \qquad (5)$$

4.2. Trapezoidal fuzzy numbers

Under universe of discourse [0,1] as described in Figure 5 has relationship $\mu_D(x)$ defined in (6)[2][3] where $x_0 \le x_1 \le x_2 \le x_3$, . $y_0 = y_3 = 0\ and\ y_1 = y_2 = 1$

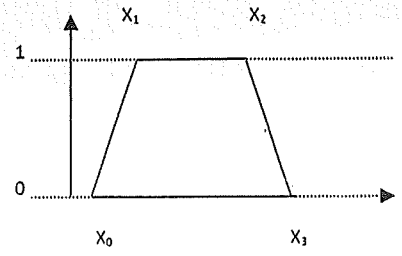


Figure 5 Trapezoidal characterized.

$$\mu_D(x)=\begin{cases}\dfrac{x-x_0}{x_1-x_0} & x_0\le x\le x_1\\ 1 & x_1\le x\le x_2\\ \dfrac{x_3-x}{x_3-x_2} & x_2\le x\le x_3\end{cases} \tag{6}$$

4.3. Pentagon fuzzy numbers

Under universe of discourse [0,1] as described in Figure 6 has relationship $\mu_P(x)$ defined in(7).

Where $x_0\le x_1\le x_2\le x_3\le x_4$, $y_0=y_4=0, y_1=y_3=0.5 and\ y_2=1$

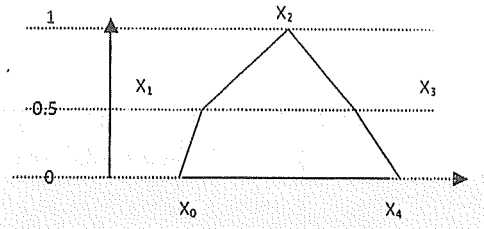


Figure 6 Pentagon characterized.

$$\mu_P(x)=\begin{cases}\dfrac{x-x_0}{x_1-x_0}y_1 & x_0\le x\le x_1\\ \dfrac{x-x_1+y_1(x_2-x)}{x_2-x_1} & x_1\le x\le x_2\\ \dfrac{x_3-x+y_3(x-x_2)}{x_3-x_2} & x_2\le x\le x_3\\ \dfrac{x_4-x}{x_4-x_3}y_3 & x_3\le x\le x_4\end{cases} \tag{7}$$

4.3. Hexagon fuzzy numbers

Under universe of discourse [0,1] as described in Figure 7 has relationship $\mu_X(x)$ defined in(8) Where $x_0\le x_1\le x_2\le x_3\le x_4\le x_5$, $y_0=y_5=0, y_1=y_4=0.5 and\ y_2=y_3=1$.

$$\mu_X(x)=\begin{cases}\dfrac{x-x_0}{x_1-x_0}y_1 & x_0\le x\le x_1\\ \dfrac{x-x_1+y_1(x_2-x)}{x_2-x_1} & x_1\le x\le x_2\\ 1 & x_2\le x\le x_3\\ \dfrac{x_4-x+y_4(x-x_3)}{x_4-x_3} & x_3\le x\le x_4\\ \dfrac{x_5-x}{x_5-x_4}y_4 & x_4\le x\le x_5\end{cases} \tag{8}$$

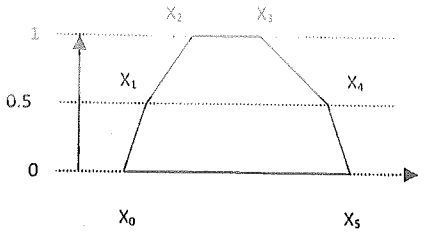


Figure 7 Hexagon characterized.

4.4. Heptagon fuzzy numbers

Under universe of discourse [0,1] as described in Figure 8 has relationship $\mu_H(x)$ defined in (9)

Where $x_0 \le x_1 \le x_2 \le x_3 \le x_4 \le x_5 \le x_6$, $y_0 = y_6 = 0$, $y_1 = y_5 = \frac{1}{3}$, $y_2 = y_4 = \frac{2}{3}$ *and* $y_3 = 1$.

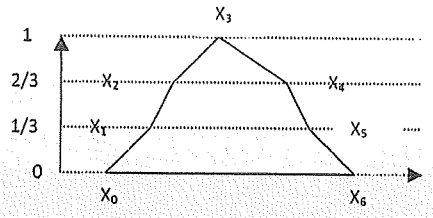


Figure 8 Heptagon characterized.

$$\mu_H(x) = \begin{cases} \frac{(x - x_0)}{(x_1 - x_0)} y_1 & x_0 \le x \le x_1 \\ \frac{(x - x_1)(y_2 - y_1)}{(x_2 - x_1)} + y_1 & x_1 \le x \le x_2 \\ \frac{(x - x_2)(1 - y_2)}{(x_3 - x_2)} + 1 & x_2 \le x \le x_3 \\ \frac{(x - x_3)(y_4 - 1)}{(x_4 - x_3)} + 1 & x_3 \le x \le x_4 \\ \frac{(x - x_4)(y_5 - y_4)}{(x_5 - x_4)} + y_6 & x_4 \le x \le x_5 \\ \frac{(x - x_6)}{(x_5 - x_6)} y_5 & x_5 \le x \le x_6 \end{cases} \quad (9)$$

4.5. Octagon fuzzy numbers

Under universe of discourse [0,1] as described in Figure 9 has relationship $\mu_O(x)$ defined in(10) Where $x_0 \le x_1 \le x_2 \le x_3 \le x_4 \le x_5 \le x_6 \le x_7$,

$y_0 = y_7 = 0, y_1 = y_6 = 1/3, y_2 = y_5 = 2/3$ *and* $y_3 = y_4 = 1$.

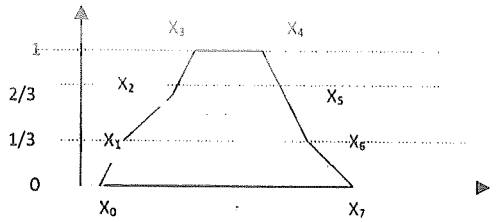


Figure 9 Octagon fuzzy shape

$$\mu_O(x) = \begin{cases} \frac{x - x_0}{x_1 - x_0} y_1 & x_0 \le x \le x_1 \\ \frac{(x - x_1)(y_2 - y_1)}{x_2 - x_1} + y_1 & x_1 \le x \le x_2 \\ \frac{(x - x_2)(1 - y_2)}{x_3 - x_2} + 1 & x_2 \le x \le x_3 \\ 1 & x_3 \le x \le x_4 \\ \frac{(x - x_4)(1 - y_3)}{x_4 - x_3} + 1 & x_4 \le x \le x_5 \\ \frac{(x - x_5)(y_4 - y_3)}{x_6 - x_5} + y_3 & x_5 \le x \le x_6 \\ \frac{x - x_7}{x_6 - x_7} y_4 & x_6 \le x \le x_7 \end{cases} \quad (10)$$

5. Numerical Example

To explain the new ranking method of criteria or alternatives, This section recalculate the center of area (centroied) of fuzzy number system which calculated before by [Chia-Chi Sun 2010], who proposed a method to evaluate the performance of top four notebook computer companies[1]. He used the TOPSIS views a MADM problem with m alternatives as a geometric system in the n-criteria. Their method is based on the concept that the chosen alternative should have the shortest distance from the positive-ideal solution and the longest distance from the negative-ideal solution [4][5]. He also extended the TOPSIS to the fuzzy environment. This method is particularly suitable for solving the group decision-making problem under fuzzy environment [7]. But he used only a triangle fuzzy number, since we can't compare their results by any fuzzy number system to check correctness. Our paper introduces many fuzzy number systems to compare between results and check correctness and accuracy. Our paper used the same inputs data and the same implementation method to employ fuzzy AHP to find the fuzzy preference weights of each criteria and alternative [6][8]. This paper modified the method of calculating by determining the weighting of evaluation criteria by introducing more polygon fuzzy number systems as trapezoidal, pentagon, hexagon, heptagon and octagon.

The steps of evaluating criteria performance weights is collected as: Constructing the AHP hierarchy which used by [Chia-Chi Sun 2010] to determining the weight of each entity related to the expert's opinions [1], Collecting the expert's opinions about sex criteria used nine linguistic values which determined by [Chia-Chi Sun 2010], who fuzzified them to triangle fuzzy numbers Table 1, But our paper fuzzy them to trapezoidal, pentagon, hexagon, heptagon and octagon fuzzy numbers to generalize the polygon fuzzy numbers Table2 Appendix A.

Table 1 Triangle fuzzy numbers

Linguistic values	Numerical scale	Triangle fuzzy numbers		
equal	1	1	1	1
weak advantage	2	1	2	3
not bad	3	2	3	4
Preferable	4	3	4	5
good	5	4	5	6
fairly good	6	5	6	7
very good	7	6	7	8
Absolutely	8	7	8	9
perfect	9	8	9	10

Using the ten pair wise matrices of six criteria each of one expert Table 3 [1].

Table 3 Experts' opinions

Expert opinion1	1	1	2	1	4	4	Expert opinion6	1	1/5	1/7	1/7	1/2	1/6
	1	1	2	1	4	4		5	1	1/7	7	7	8
	1/2	1/2	1	1/2	2	2		7	7	1	8	9	8
	1	1	2	1	4	4		7	1/7	1/8	1	8	5
	1/4	1/4	1/2	1/4	1	1		2	1/7	1/9	1/8	1	1/6
	1/4	1/4	1/2	1/4	1	1		6	1/8	1/8	1/5	6	1
Expert opinion2	1	1	4	1	3	3	Expert opinion7	1	1/9	1	1	9	1
	1	1	4	1	3	3		9	1	1	9	9	9
	1/4	1/4	1	1/4	1/2	1/3		1	1	1	9	9	9
	1	1	4	1	3	2		1	1/9	1/9	1	2	2
	1/3	1/3	2	1/3	1	1		1/9	1/9	1/9	1/2	1	1
	1/3	1/3	3	1/2	1	1		1	1/9	1/9	1/2	1	1
Expert opinion 3	1	2	3	3	2	4	Expert opinion 8	1	3	4	1	6	1/2
	1/2	1	6	4	2	1		1/3	1	2	1	6	4
	1/3	1/6	1	1	3	4		1/4	1/2	1	3	2	5
	1/3	1/4	1	1	3	4		1	1	1/3	1	5	1
	1/2	1/2	1/3	1/3	1	3		1/6	1/6	1/2	1/5	1	1
	1/4	1	1/4	1/4	1/3	1		2	1/4	1/5	1	1	1
Expert opinion 4	1	3	4	1	6	1	Expert opinion 9	1	3	2	1	1	2
	1/3	1	1	1	1	1		1/3	1	4	5	1	3
	1/4	1	1	1	2	2		1/2	1/4	1	5	3	1
	1	1	1	1	1	1		1	1/5	1/5	1	2	1
	1/6	1	1/2	1	1	2		1	1	1/3	1/2	1	3
	1	1	1/2	1	1/2	1		1/2	1/3	1	1	1/3	1
Expert opinion 5	1	3	1	1	2	3	Expert opinion 10	1	3	1	2	6	1
	1/3	1	3	1	5	3		1/3	1	2	3	4	2
	1	1/3	1	1	3	1		1	1/2	1	5	1	1
	1	1	1	1	1	2		1/2	1/3	1/5	1	3	5
	1/2	1/5	1/3	1	1	3		1/6	1/4	1	1/3	1	1
	1/3	1/3	1	1/2	1/3	1		1	1/2	1	1/5	1	1

Convert each pair wise comparison value to polygon fuzzy number in tables 1,2 and calculate the fuzzy geometric mean of each criteria as defined in (11) and represent the results of triangle, trapezoidal, heptagon, hexagon, octagon and octagon in Tables 4, 5, 6, 7, 8 and 9 respectively.

$$\tilde{r_i} = (\tilde{a_{i1}} \otimes \tilde{a_{i2}} \otimes \ldots\ldots \otimes \tilde{a_{in}})^{\frac{1}{n}} \qquad (11)$$

where $\tilde{a_{ij}}$ is a fuzzy pair wise comparison between i and j criterion on decision makers matrices.

Table 4 Triangle fuzzy geometric mean

r_1	1.143	1.416	1.668
r_2	1.466	1.786	2.105
r_3	1.006	1.235	1.530
r_4	0.894	1.055	1.213
r_5	0.427	0.518	0.655
r_6	0.500	0.586	0.728

Table 5 Trapezoidal fuzzy geometric mean

r_1	1.143	1.286	1.541	1.668
r_2	1.534	1.703	2.016	2.174
r_3	1.006	1.119	1.366	1.530
r_4	0.894	0.978	1.132	1.213
r_5	0.467	0.513	0.622	0.699
r_6	0.500	0.539	0.646	0.728

Table 6 Pentagon fuzzy geometric mean

r_1	1.143	1.286	1.416	1.541	1.668
r_2	1.466	1.631	1.786	1.941	2.129
r_3	1.006	1.119	1.235	1.366	1.548
r_4	0.894	0.978	1.055	1.132	1.213
r_5	0.427	0.470	0.518	0.576	0.663
r_6	0.500	0.539	0.586	0.646	0.753

Table 7 Hexagon fuzzy geometric mean

r_1	1.143	1.217	1.238	1.541	1.603	1.668
r_2	1.466	1.551	1.631	1.941	2.021	2.105
r_3	1.006	1.063	1.119	1.366	1.442	1.530
r_4	0.894	0.937	0.907	1.132	1.172	1.213
r_5	0.427	0.448	0.404	0.577	0.612	0.655
r_6	0.500	0.519	0.500	0.646	0.683	0.728

Table 8 Heptagon fuzzy geometric mean

r_1	1.143	1.217	1.286	1.416	1.541	1.603	1.668
r_2	1.466	1.551	1.631	1.786	1.941	2.022	2.105
r_3	1.006	1.451	1.119	1.235	1.366	1.443	1.540
r_4	0.894	0.987	0.978	1.055	1.132	1.172	1.213
r_5	0.427	0.552	0.470	0.518	0.577	0.612	0.655
r_6	0.497	0.743	0.536	0.582	0.640	0.676	0.719

Table 9 Octagon fuzzy geometric mean

r_1	1.1433	1.2169	1.2857	1.3516	1.4782	1.5407	1.6032	1.6679
r_2	1.4663	1.5505	1.6307	1.709	1.8633	1.9413	2.0212	2.1049
r_3	1.0061	1.063	1.1185	1.1753	1.2979	1.3663	1.4422	1.5296
r_4	0.8936	0.9371	0.9776	1.0167	1.0937	1.1324	1.1718	1.2135
r_5	0.3082	0.3221	0.3367	0.3525	0.3902	0.4131	0.4391	0.4716
r_6	0.4998	0.5192	0.5394	0.5614	0.614	0.6457	0.6827	0.7277

Calculate the performance fuzzy weight of each criteria as defined in (12) and illustrate the results in Tables 10, 11, 12, 13, 14 and 15 to represent triangle, trapezoidal, pentagon, hexagon, heptagon and octagon fuzzy performance weight of each criteria.

$$\tilde{w_i} = \tilde{r_i} \otimes (\tilde{r_1} \oplus \tilde{r_2} \oplus \ldots\ldots \oplus \tilde{r_n})^{-\frac{1}{n}} \qquad (12)$$

Where $\tilde{r_i}$ is fuzzy geometric mean of each criteria.

Table 10 Fuzzy performance triangle weight

w_1	0.145	0.215	0.307
w_2	0.186	0.271	0.387
w_3	0.127	0.187	0.281
w_4	0.113	0.160	0.223
w_5	0.054	0.078	0.121
w_6	0.063	0.089	0.134

Table 11 Fuzzy performance trapezoidal weight

w_1	0.143	0.176	0.251	0.301
w_2	0.191	0.233	0.329	0.392
w_3	0.126	0.153	0.223	0.276
w_4	0.112	0.134	0.184	0.219
w_5	0.058	0.070	0.101	0.126
w_6	0.062	0.074	0.105	0.131

Table 12 Fuzzy performance pentagon weight

w_1	0.143	0.179	0.215	0.256	0.307
w_2	0.184	0.226	0.271	0.322	0.392
w_3	0.126	0.155	0.187	0.227	0.285
w_4	0.112	0.136	0.160	0.188	0.223
w_5	0.054	0.065	0.078	0.096	0.122
w_6	0.063	0.075	0.089	0.107	0.139

Table 13 Fuzzy performance hexagon weight

w_1	0.145	0.162	0.172	0.266	0.280	0.307
w_2	0.186	0.206	0.226	0.335	0.352	0.387
w_3	0.127	0.141	0.155	0.236	0.251	0.281
w_4	0.113	0.124	0.126	0.195	0.204	0.223
w_5	0.054	0.060	0.056	0.099	0.107	0.121
w_6	0.063	0.069	0.069	0.111	0.119	0.134

Table 14 Fuzzy performance heptagon weight

w_1	0.145	0.162	0.179	0.215	0.256	0.247	0.307
w_2	0.186	0.206	0.227	0.271	0.323	0.311	0.387
w_3	0.127	0.193	0.155	0.187	0.227	0.222	0.283
w_4	0.113	0.131	0.136	0.160	0.188	0.180	0.223
w_5	0.054	0.073	0.065	0.079	0.096	0.094	0.121
w_6	0.063	0.099	0.075	0.088	0.106	0.104	0.132

Table 15 Fuzzy performance octagon weight

w_1	0.1482	0.1653	0.1826	0.2006	0.2397	0.2616	0.2858	0.3137
w_2	0.1901	0.2107	0.2317	0.2537	0.3022	0.3297	0.3604	0.3959
w_3	0.1304	0.1444	0.1589	0.1745	0.2105	0.232	0.2571	0.2877
w_4	0.1158	0.1273	0.1389	0.1509	0.1774	0.1923	0.2089	0.2282
w_5	0.0399	0.0438	0.0478	0.0523	0.0633	0.0702	0.0783	0.0887
w_6	0.0648	0.0705	0.0766	0.0833	0.0996	0.1097	0.1217	0.1369

Defuzzyficate the weights values of each polygon by calculate the center of area of each one defined in (13), (14), (15), (16), (17) and (18), then represent the values in Table 16for triangle, trapezoidal, pentagon, hexagon, heptagon and octagon.

$$C_T = [(x_2 - x_0) + (x_1 - x_0)]/3 + x_0 \tag{13}$$

$$C_D = \frac{x_2^2 - x_0^2 + x_3^2 - x_1^2 + x_2 x_3 - x_0 x_1}{3*((x_2 + x_3) - (x_0 + x_1))} \tag{14}$$

$$C_P = \frac{x_4^2 - x_0^2 + 2x_2^2 - 2x_1^2 + x_2 x_4 - x_3 x_1 + x_3 x_2 - x_0 x_1}{3*((x_4 + 2x_2) - (x_0 + 2x_1))} \tag{15}$$

$$C_X = \frac{x_5^2 - x_0^2 + 2x_3^2 + x_2^2 - x_4^2 - 2x_1^2 + x_2 x_5 - x_4 x_1 + x_3 x_5 - x_0 x_1}{3*((x_5 + 2x_3 + x_2) - (x_0 + 2x_1 + x_4))} \tag{16}$$

$$C_H = \frac{x_6^2 + 2x_5^2 + 2x_4^2 + x_2^2 - 2x_1^2 - x_0^2 + x_6 x_5 + x_5 x_4 + x_4 x_3 - x_3 x_2 + 2x_2 x_1 - x_1 x_0}{3(x_6 + 2x_5 + 2x_4 - 2x_2 - 2x_1 - x_0)} \tag{17}$$

$$C_O = \frac{x_7^2 + 2x_6^2 + 2x_5^2 + 13x_4^2 - x_3^2 + 2x_2^2 - 2x_1^2 - x_0^2 + x_7 x_6 + x_6 x_5 + 13x_5 x_4 - x_4 x_3 - x_3 x_2 - x_2 x_1 - x_1 x_0}{3(x_7 + 4x_6 + 6x_5 + 3x_4 - x_3 - 2x_2 - 2x_1 - x_0)} \tag{18}$$

Table 16 Defuzzyfication and Ranking

Triangle Ranking		Trapezoidal Ranking		Pentagon Ranking		Hexagon Ranking		Heptagon Ranking		Octagon Ranking	
0.222	2	0.218	2	0.259	2	0.255	2	0.339	2	0.2284	2
0.281	1	0.287	1	0.322	1	0.326	1	0.434	1	0.2881	1
0.199	3	0.195	3	0.230	3	0.213	3	0.352	3	0.2001	3
0.165	4	0.162	4	0.201	4	0.222	4	0.277	4	0.1697	4
0.084	6	0.089	6	0.101	6	0.098	6	0.137	6	0.0602	6
0.095	5	0.094	5	0.115	5	0.116	5	0.186	5	0.0949	5

We note that the six fuzzy polygons have the same ranking of alternatives which they are the same ranking of the triangle fuzzy number of the example in pervious puplished, This improves that, the new method has the same quality and efficiency of [Chia-Chi Sun 2010] [1], But the proposed method introduced more flexibility, availability and accuracy of results and the ability of comprising between results to check correctness and precise.

6. Conclusion

This paper introduced first the level set lemma which distributes the polygon to a number of heights corresponding to the number of their vertices, this paper also introduced an improved theorem to calculate the centre of area of any polygon to help in ranking the criteria or alternatives of fuzzy multi criteria decision making problems. From the proposed theorem, we deduced definitions of triangle, trapezoidal, pentagon and hexagon, heptagon and octagon fuzzy numbers, since the deductive definitions of triangle and trapezoidal are coincide to the corresponding definitions on previous researches, which insure the correctness of the proposed method. That is meaning that pentagon, hexagon, heptagon and octagon definitions are correct and accurate. This paper also solved an example by the six fuzzy number systems, which have been solved before by a triangle fuzzy number only. This paper can compare between results of the different fuzzy number systems to test correctness and accuracy, where it can be difficult to compare between results before. The congruencies of different fuzzy number results improve the correctness of the theorem and the correctness of the deductive definition. This paper leads us to the generalized polygon fuzzy number systems which can solve any fuzzy multi criteria decision making problem with any number of vertices, this paper introduces the flexibility to solve any fuzzy multi criteria decision making problem with different fuzzy numbers.

References

[1] Chia-Chi Sun, Institute of Technology Management, National Chiao Tung University, No. 1001, Ta-Hsuch Rd., Hsinchu 300, Taiwan, A performance evaluation model by integrating fuzzy AHP and fuzzy TOPSIS methods, Expert Systems with Applications 37 (2010) 7745–7754.

[2] Mohammad Anisseh · Fatemeh Piri · Mohammad Reza Shahraki ·Fazlollah Agamohamadi, Fuzzy extension of TOPSIS model for group decision making under multiple criteria, Springer Science+Business Media B.V. 2011, DOI 10.1007/s10462-011-9258-2.

[3] E. Roghanian *, J. Rahimi, A. Ansari, Department of Industrial Engineering, K.N Toosi University of Technology, Tehran, Iran, Comparison of first aggregation and last aggregation in fuzzy group TOPSIS, Applied Mathematical Modelling 34 (2010) 3754–3766.

[4] T. Aouam, S.I. Chang *, E.S. Lee, Fuzzy MADM: An outranking method, Department of Industrial and Manufacturing Systems Engineering, Kansas State University, Manhattan, KS 66506, USA, European Journal of

Operational Research 145 (2003) 317–328.

[5] J.-R. Chang · T.-H. Ho · C.-H. Cheng · A.-P. Chen , Dynamic fuzzy OWA model for group multiple criteria decision MAKING, Springer-Verlag, DOI 10.1007/s00500-005-0484-x, (2006) 10: 543–554.

[6] L.C. Leung a, D. Cao, Decision Sciences and Managerial Economics, Faculty of Business Administration, The Chinese University of Hong Kong, Hong Kong, People's Republic of Chinab School of Management, Shenzhen University, Guangdong, People's Republic of China, Theory and Methodology On consistency and ranking of alternatives in fuzzy AHP, European Journal of Operational Research 124 (2000) 102±113.

[7] Mehmet Kabak a, Serhat Burmaog˘lu a, Yig it Kazancoglu, Turkish Army Academy, System Management Sciences Department, Bakanliklar, Ankara, Turkey b Izmir University of Economics, Turkey, A fuzzy hybrid MCDM approach for professional selection, Expert Systems with Applications 39 (2012) 3516–3525.

[8] Yu-Jie Wanga,_, Hsuan-Shih Leeb, Department of International Trade, Lan Yang Institute of Technology, I Lan 261, Taiwan, ROC b Department of Shipping and Transportation Management, National Taiwan Ocean University, Keelung 202, Taiwan, ROC, Generalizing TOPSIS for fuzzy multiple-criteria group decision-making, Computers and Mathematics with Applications 53 (2007) 1762–1772.

[9] Hao Fanghua· Chen Guanchun, A Fuzzy Multi-Criteria Group Decision-Making Model Based on Weighted Borda Scoring Method for Watershed Ecological Risk Management: a Case Study of Three Gorges Reservoir Area of China, Springer Science and Business Media B.V. 2009, DOI 10.1007/s11269-009-9544-9, (2010)24:2139–2165.

Appendix A

Table 2 Polygon fuzzy numbers

N. S.	Trapezoidal fuzzy numbers				Pentagon fuzzy numbers					Hexagon fuzzy numbers						Heptagon fuzzy numbers							Octagon fuzzy numbers							
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1.5	2.5	3	1	1.5	2	2.5	3	1	1.25	1.5	2.5	2.75	3	1	1.25	1.5	2	2.5	2.75	3	1	1.25	1.5	1.75	2.25	2.5	2.75	3
3	2	2.5	3.5	4	2	2.5	3	3.5	4	2	2.25	2.5	3.5	3.75	4	2	2.25	2.5	3	3.5	3.75	4	2	2.25	2.5	2.75	3.25	3.5	3.75	4
4	3	3.5	4.5	5	3	3.5	4	4.5	5	3	3.25	3.5	4.5	4.75	5	3	3.25	3.5	4	4.5	4.75	5	3	3.25	3.5	3.75	4.25	4.5	4.75	5
5	4	4.5	5.5	6	4	4.5	5	5.5	6	4	4.25	4.5	5.5	5.75	6	4	4.25	4.5	5	5.5	5.75	6	4	4.25	4.5	4.75	5.25	5.5	5.75	6
6	5	5.5	6.5	7	5	5.5	6	6.5	7	5	5.25	5.5	6.5	6.75	7	5	5.25	5.5	6	6.5	6.75	7	5	5.25	5.5	5.75	5.25	6.5	6.75	7
7	6	6.5	7.5	8	6	6.5	7	7.5	8	6	6.25	6.5	7.5	7.75	8	6	6.25	6.5	7	7.5	7.85	8	6	6.25	6.5	6.75	6.25	7.5	7.75	8
8	7	7.5	8.5	9	7	7.5	8	8.5	9	7	7.25	7.5	8.5	8.75	9	7	7.25	7.5	8	8.5	8.75	9	7	7.25	7.5	7.75	7.25	8.5	8.75	9
9	8	8.5	9.5	10	8	8.5	9	9.5	10	8	8.25	8.5	9.5	9.75	10	8	8.25	8.5	9	9.5	9.75	10	8	8.25	8.5	8.75	8.25	9.5	9.75	10

An Experiment for Performance Evaluation of Peer-to-Peer Mobile Web Services

Zaki A. M. [1]

Abstract

Mobile Web Service (MWS) is the implementation of Services Oriented Architecture (SOA) over Mobile Devices. Many technologies enable services to be consumed from mobile device. Increased connectivity of mobile devices and enhancements of mobile hardware capabilities are the main motive to ask the questions: can mobile device be mobile server? Can mobile device be a service provider? What about performance of web services provided by mobile device and consumed by another?

In this paper, an experiment for a performance evaluation of peer-to-peer mobile web services will be demonstrated. This experiment is conducted using precompiled lightweight web server (lighttpd), PHP and MySQL database server for ARM Architecture. The experiment uses Android Emulator to evaluate performance of Simple Object Access Protocol (SOAP) based Services in no Web Service Description Language (no-WSDL) mode versus Representational State Transfer (REST) based Services. The experiments show that SOAP based services in no-WSDL mode perform faster than REST based Services in terms of Response Time.

Keywords

Mobile Computing, Web Services, Mobile Web Services, SOAP, REST.

I. Introduction

Web services (WS) are a machine-to-machine communication standard that used to build distributed application over the internet [1]. There are two flavors of web services, Simple Object Access Protocol- Based (SOAP-Based WS) and Representation State-full Transfer –based (REST-Based WS) [1]. Both of them use Hypertext Transfer Protocol (HTTP) to communicate. However, the major difference that SOAP uses HTTP to send SOAP envelops which used in communication. Where REST Based Services uses HTTP to create, read, update and delete Service Resource [2]. Web services can be consumed by mobile application. In addition, mobile devices can be a service provider [2] [3]. This introduces the basis of distributed mobile applications.

[1] *CIS Teaching Assistant, Sadat academy for Management Sciences , a.m.zaki@sadatacademy.edu.eg*

Web service flavor that is based over SOAP can be used in two modes. First mode is WSDL Mode, in which service is descried in xml and this WSDL file published over service registry. Then service consumer search for it and then binds to service provider and consume the service. The other mode is no-WSDL mode, in which the service consumer already knows the service specifications and location and manually binds to service provider and consume the service [4]. Web services based on REST uses the idea that services are resources. These resources used by calling it using its Unformed Resource Locator (URL). REST based uses HTTP Methods to work with resources Get to retrieve a resource, Post to Create a resource, Put to edit a resource and Delete to remove a resource [5].

This paper is organized as the following, Section (II) review related researches. Section (III) describe experiment environment .Section (IV) describes Services used in experiment. Section (V) shows the experiment results. Finally, section (VI) contains conclusions and future work.

II. Related Researches

Mobile web service is a hot topic of research. Many researches since 2003 have been done. These researches were in the following areas: mobile web service provisioning and performance of mobile web services. In the area of mobile web service provisioning, many frameworks were introduced e.g. Mobile Host framework [6] , Handheld Flexible Representation (HHFR) [7], and a simple implementation for a mobile web service [8]. In 2004, a research conducted to evaluate mobile web service with compressed messages [9]. In 2005, a research conducted to evaluate performance of HHFR framework [7] . In addition , performance evaluation of web services on embedded devices was conducted [10].In 2009, a comparative study between Mobile web services over Wireless Sensor, Stationary Machine, Mobile Device and note book was conducted [11]. In 2010, another performance evaluation for mobile web service was conducted [2] where mobile device used as service consumer. Finally, in 2011, another experiment was conducted [3]. All of these researches show that it is possible to implement web services over mobile device. It also performs a comparison between different web services architectures.

III. Experiment Environment

This experiment was conducted using android mobile emulator. Using two nodes both of them contain a lighttpd [12] server, PHP Interpreter and MySQL server. Also In this research, RESTLER framework used to enable

REST based services over mobile devices [13] and PHP-SOAP for SOAP based Services [4].

Node Configurations

Table 1 Nodes Configuration

Nodes	Node 1 (Nexus 7)	Node 2 (Nexus 4)
CPU – ABI	ARM (armabi-v7a) , Cortex-a8	
Mobile System Image	Android 4.2.2 (API 17)	
RAM	512 MB	343 MB
SD-Card	1024 MB	
Internal Storage	200	
Http Server	lighttpd 1.4	
Scripting Language	PHP	

Nodes Layout

In peer-to-peer model, two nodes communicate over ad-hoc network. Each node has the following software as shown in the following figure:

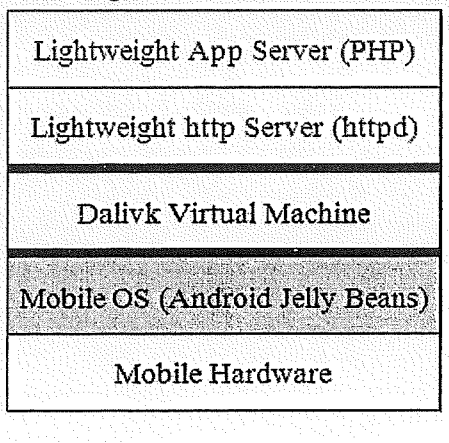


Figure 1 Node Software Stack

The two nodes communicate over hosting machine Because of network isolation of each node instance. The following figure shows the communication between peers [14]

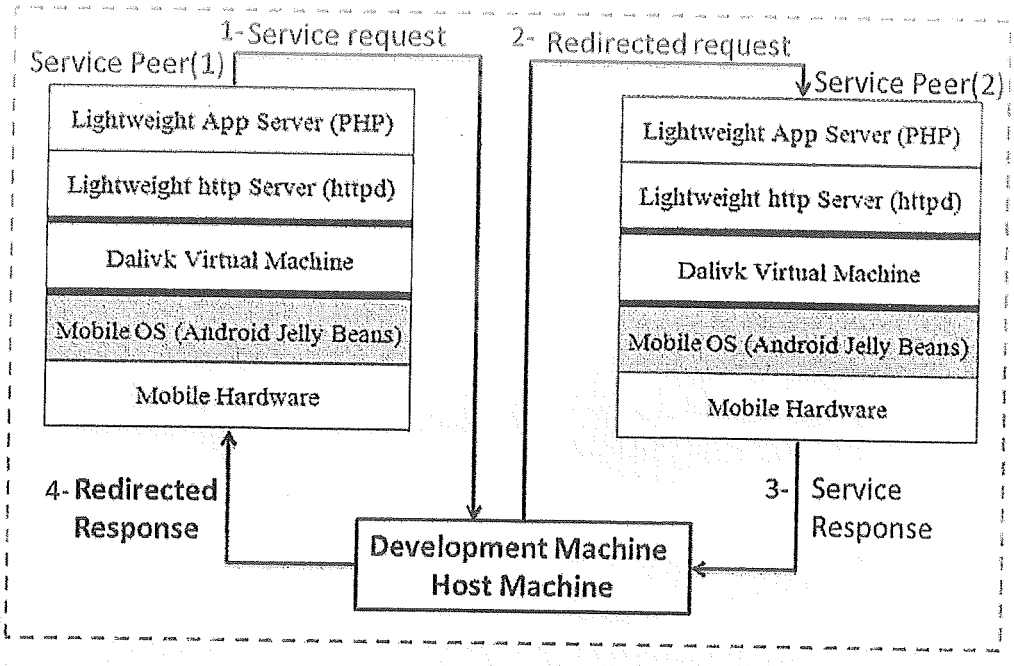


Figure 2 Nodes Layout and Communication Model

Hosting Machine Configuration

The hosting Environment configuration was:

Table 2 Hosting Machine Configuration

CPU	4x Intel(R) Core(TM) i5-2430M CPU @ 2.40GHz
Memory	8088MB
OS	Ubuntu 12.04.3 LTS, Kernel: Linux 3.11.2-031102-generic (x86_64)
Android SDK	Revision No 22.2.1

IV. Service Benchmarking

Concatenation Services

In this experiment, service used in benchmarking is concatenation service. Service consumer sends an array of characters with n bytes size and service provider concatenates the message with another string and sends it back to consumer.

Image Transfer

In this experiment, service used in benchmarking is Image Transfer service. Service consumer sends a request for an image and service sends it back to consumer.

V. Experiment Results

SOAP based concatenation Service:

Table 3 SOAP-Based Concatenation Service

Message Size in Bytes	Average Response Time (Millisecond)	Standard Deviation	Coefficient Of Variations
5	*14.493*	*6.093*	*42.041%*
10	*15.013*	*5.478*	*36.485%*
15	*14.860*	*5.571*	*37.488%*
20	*14.686*	*5.275*	*35.921%*
25	*15.622*	*6.358*	*40.698%*
30	*13.744*	*6.311*	*45.915%*
35	*13.388*	*4.059*	*30.319%*
40	*15.338*	*5.076*	*30.319%*

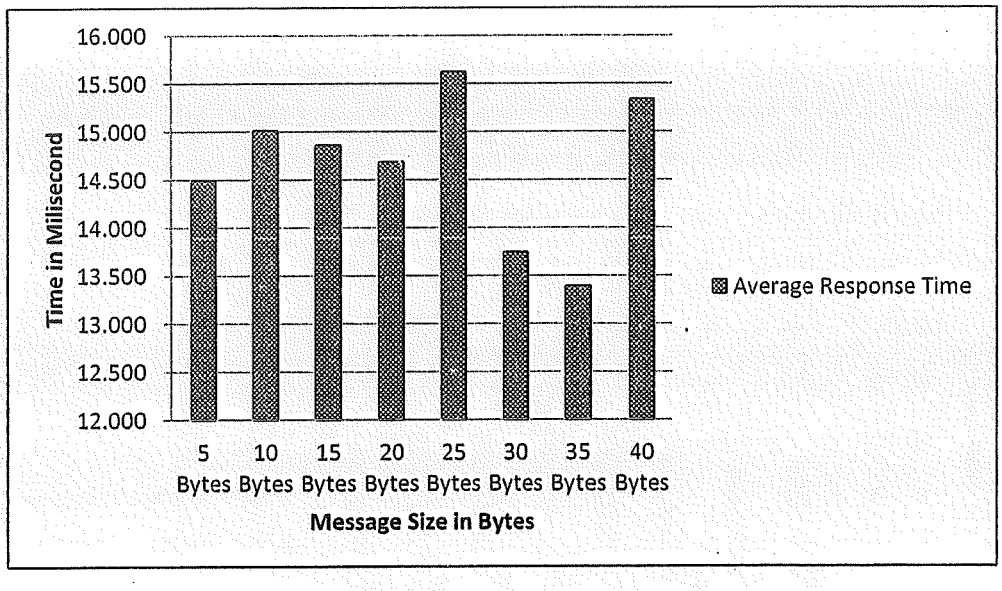


Figure 3 Average Response Time of SOAP-Based Concatenation Service

REST based concatenation Service:

Table 4 REST Based Concatenation Service

Message Size in Bytes	Average Response Time (Millisecond)	Standard Deviation	Coefficient Of Variations
5	53.730	13.316	24.783%
10	49.241	8.076	16.401%
15	62.387	27.941	44.787%
20	66.753	21.870	32.762%
25	76.659	40.719	53.117%
30	64.899	14.517	22.369%
35	64.258	11.048	17.193%

Message Size in Bytes	Average Response Time (Millisecond)	Standard Deviation	Coefficient Of Variations
40	64.315	10.424	16.208%

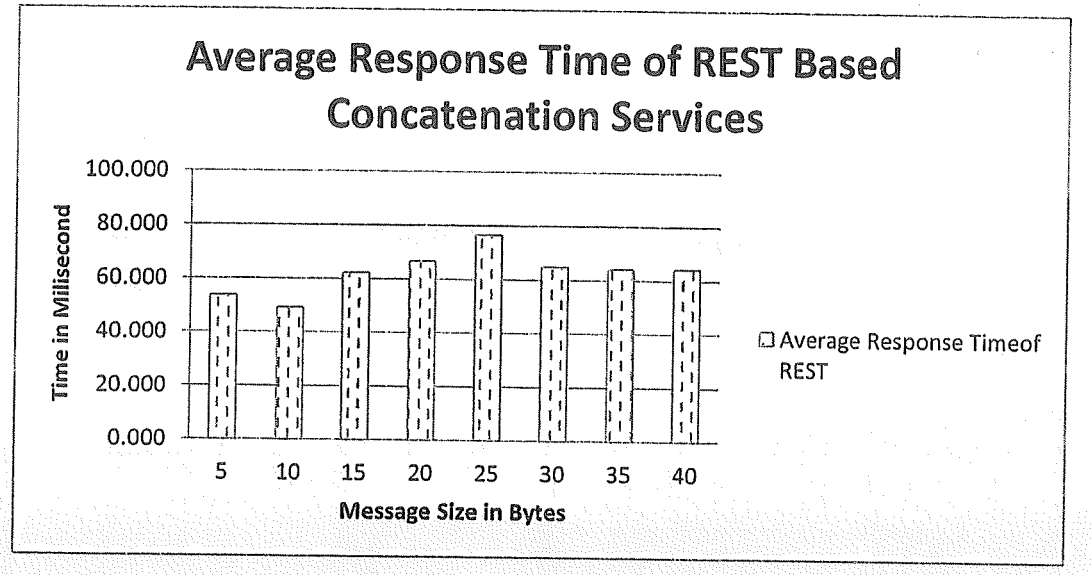


Figure 4 Average Response Time of REST Based Concatenation Service

SOAP based Vs. REST based concatenation Service:

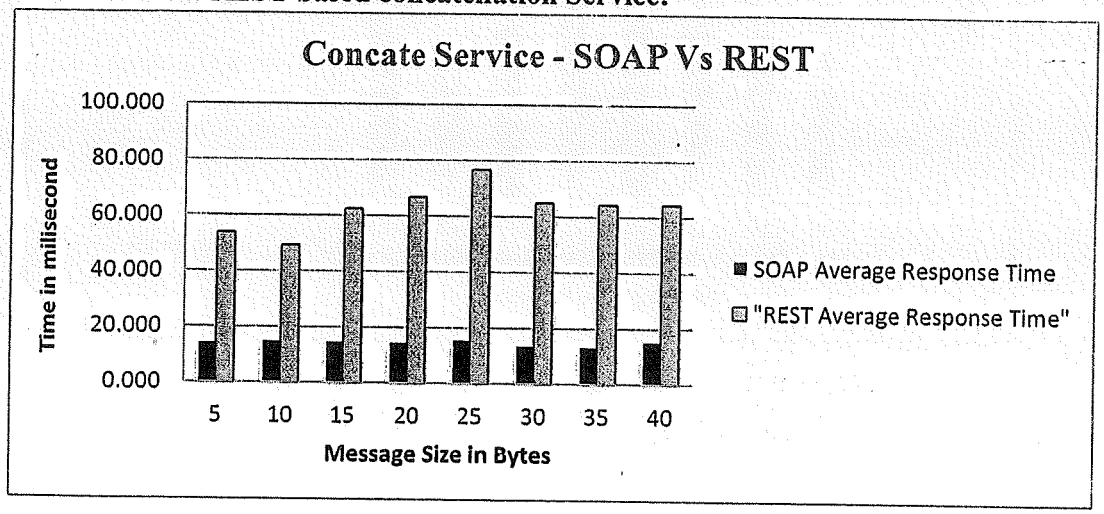


Figure 5 SOAP Based Vs. REST Based Concatenation Service

SOAP based image transfer:

Table 5 SOAP Based Image Transfer

Image Resolution	Average Response Time (Millisecond)	Standard Deviation	Coefficient Of Variations
32 × 38	14.365	4.590	31.95%
64 × 76	15.043	5.271	35.04%
128 × 152	16.684	4.857	29.11%
256 × 304	21.906	6.431	29.36%
512 × 608	43.694	10.874	24.89%

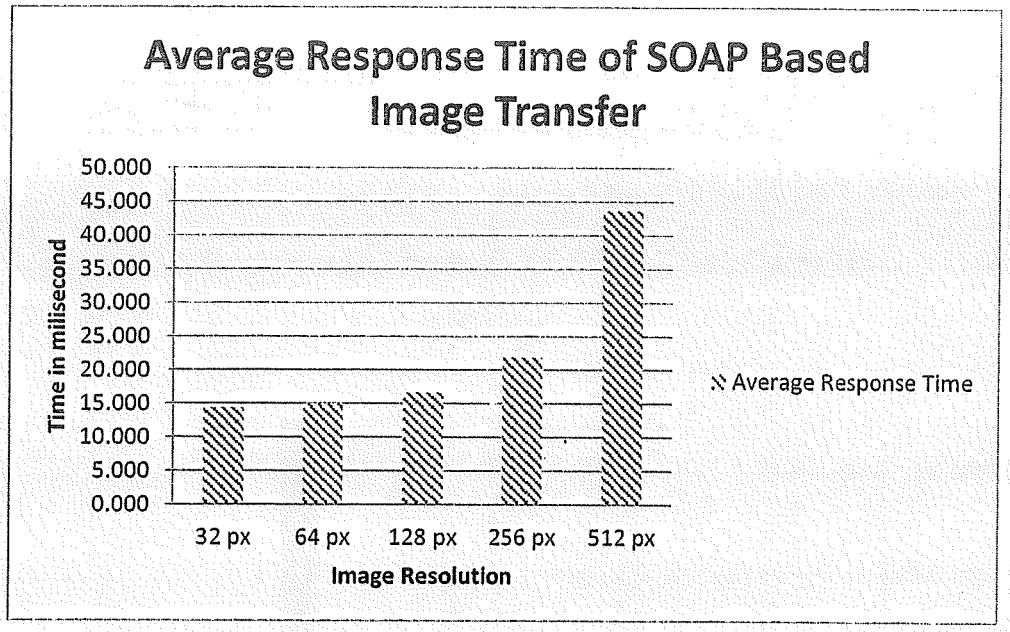


Figure 6 Average Response Time of SOAP Based Image Transfer

REST based image transfer:

Table 6 REST based Image Transfer

Image Resolution	Average Response Time (Millisecond)	Standard Deviation	Coefficient Of Variations
32 × 38	66.930	20.080	30.00%
64 × 76	81.850	17.466	21.34%
128 × 152	148.361	26.443	17.82%
256 × 304	507.282	126.316	24.90%
512 × 608	2285.936	804.554	35.20%

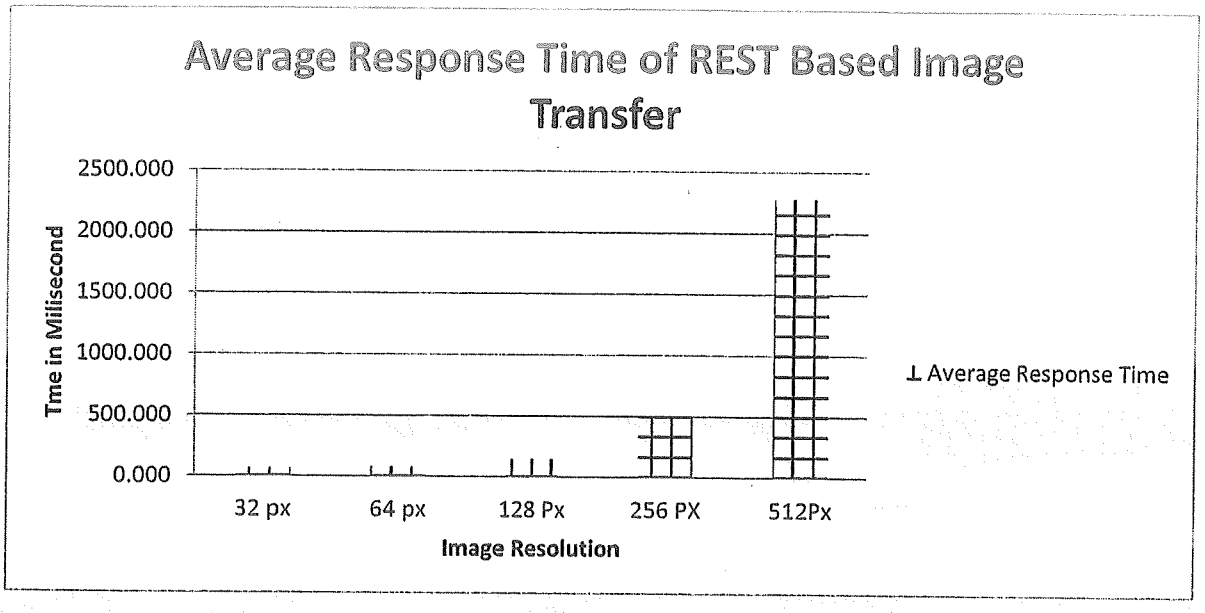


Figure 7 Average Response Time of REST based Image Transfer

SOAP based Vs. REST based image transfer:

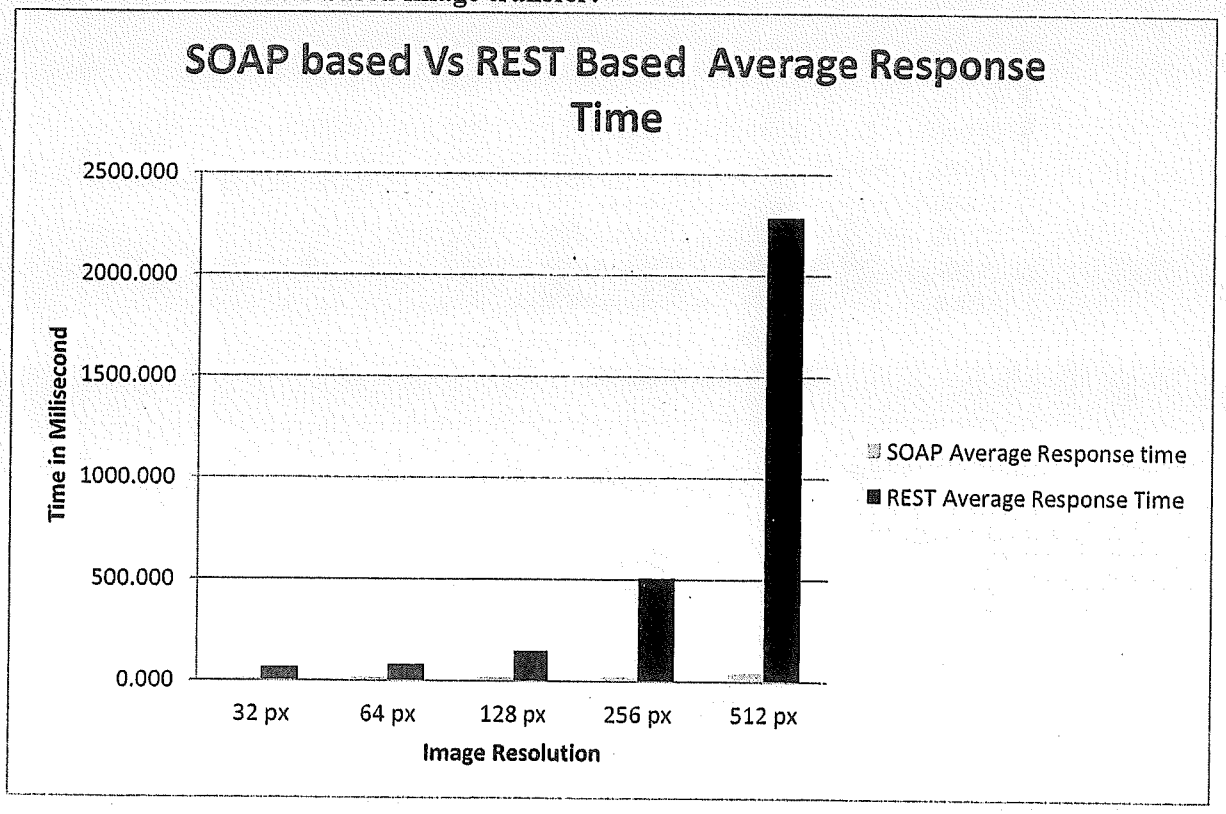


Figure 8 SOAP Vs. REST Based Image Transfer

VI. Conclusions and future work

From previous experiment, we can conclude:

- The no-WSDL SOAP based services responds in less time rather than REST based service in the two experiments
- REST Based services are easier to implement rather than SOAP based services.
- Mobile servers fail to handle concurrent requests.
- To save power, the mobile device sleeps after a specific duration, which leads to server failure.
- Mobile phone devices fit more as a thin client.
- Running Servers over mobile devices is power consuming.
- Message Size is not the only factor of response time. Many factors can affect the services response like network status, location, and running application in background.

As a future work:

- Android OS can be used over sensors, but needs modifications in user interface and OS Services.
- This experiment needs to be done over real devices to test it over different mobile networks.
- Test Database Services on MySQL and SQLite.
- Test another types of images.

VII. References

[1] David Booth, Hugo Haas, Francis McCabe,Eric Newcomer, Michael Champion,Chris Ferris,David Orchard. (2003, August) w3.org. [Online]. http://www.w3.org/TR/WD-ws-arch-20030808/#id2617948

[2] Hatem Hamad, Motaz Saad, and Ramzi Abed, "Performance Evaluation of RESTful Web Services for Mobile Devices," *International Arab Journal of e-Technolog,* vol. 1, January 2010.

[3] Rabeb Mizouni , M. Adel Serhani , Rachida Dssouli, "Performance Evaluation of Mobile Web Services," in *Ninth IEEE European Conference on Web Services*, 2011.

[4] PHP Manual. [Online]. http://php.net/manual/en/book.soap.php

[5] Alex Rodriguez, "RESTful Web services: The basics," 2008.

[6] Satish Narayana Srirama, Matthias Jarke, Wolfgang Prinz, "Mobile Web Service Provisioning," in *AICT-ICIW '06*, Aachen, 2006.

[7] Sangyoon Oh, Geoffrey C. Fox, "HHFR: A new architecture for Mobile Web Services Principles and Implementations," 2005.

[8] Stefan Berger, Scott McFaddin, Chandra Narayanaswami, Mandayam Raghunath, "Web Services on Mobile Devices – Implementation and Experience," , 2003.

[9] M. Tian, T. Voigt, T. Naumowicz, H. Ritter, J. Schiller, "Performance Considerations for Mobile Web Services," 2004.

[10] Daniel Schall , Marco Aiello , Schahram Dustdar, "Web Services on Embedded Devices," *J. WEB INFOR. SYST*, 2005.

[11] Winfried Lamersdorf , Sonja Zaplata , Viktor Dreiling, "Realizing Mobile Web Services for Dynamic Applications," *AIS Transactions on Enterprise Systems*, 2009.

[12] Lighttpd Fly Light. [Online]. http://www.lighttpd.net/

[13] (2007) http://luracast.com/. [Online]. http://luracast.com/products/restler

[14] (2013) Exploring the SDK - Android Developers. [Online]. http://developer.android.com/sdk/exploring.html

An Enhanced Term Weighting Algorithm: A Fuzzy Ontology Approach

Zeinab E. Attia*, Ahmed M. Gadallah*, Hesham A. Hefny*

Abstract

This paper proposes a term weighting algorithm for research papers. The proposed algorithm weights a paper's annotated keywords according to a certain view. Generally it uses a predefined multi-view fuzzy ontology for detecting related concepts and terms to each keyword. Accordingly, the annotation process will be achieved respecting all related concepts and terms to each keyword. The proposed algorithm is tested and compared with an ontology based term weighting algorithm, namely, Fernandez term weighting algorithm. The tests show that it enhances the resulted weights accuracy.

Keywords: Term weighting, Fuzzy ontology, Document annotation

1. Introduction

An information retrieval system (IR) consists of a set of documents, a user query, a retrieval engine and a ranking module. It stores and indexes documents, such that when users express their information needs in a query, the retrieval engine retrieves a set of relevant documents associating a score to each one. The higher the score is, the greater the document relevance. Then the ranking module ranks them and displays them to the user [4]. In order to index a set of documents an annotation algorithm should be used.

Annotation is a process of adding metadata to a certain document in order to describe it. This metadata may include a string of weighted keywords, authors' names, the publishing conference or journal, date of publishing ...etc. The weight of each keyword reflects to what extend does it represents that document. The process of weighting a document's keywords is known as *Term Weighting* (TW).

The paper presents a domain specific Term Weighting approach based on a multi view fuzzy ontology. The paper is organized as follows: the next section presents term weighting techniques. A survey about fuzzy ontology is conducted in section 3. Section 4 presents some related works. The proposed term weighting algorithm is presented in section 5. The proposed algorithm is discussed and concluded in Section 6.

* Department of Computer and information sciences, Institute of statistical studies and research, Cairo University

2. Term Weighting

Term weighting (TW) is the process of calculating a weight for each term representing a certain document. Such weight reflects to what extent this term represents its document. Due to its importance, term weighting is used in many fields such as document clustering, information retrieval (IR), and many more. Regarding IR, it enhances both of recall and precision measures. Also it enhances the rank of the retrieved documents [1].

There are several algorithms representing the term weighting process. Such algorithms are applied either on a domain specific or on a general one. Almost, all general term weighting algorithms are statistics-based algorithms. One of the most popular term weighting algorithms for the general Term Weighting approach is the Term Frequency Inverse Document Frequency, *TFIDF*. *TFIDF* is a statistically based algorithm that uses the following equation [2]:

$$TFIDF_{ij}=TF_{ij}*IDF_i \quad (1)$$

$$TF_{ij}=\frac{the\ number\ of\ occurening\ the\ term\ Ti\ in\ the\ document\ dj}{the\ number\ of\ all\ terms\ in\ the\ document\ dj} \quad (2)$$

$$IDF_i=\log\frac{N}{n} \quad (3)$$

Where, TF_{ij} is the frequency of occurring the term t_i in the document d_j with respect to the number of all words in that document. IDF is the inverse document frequency. N is the number of documents in the collection. n is the number of documents that contain the term t_i.

Domain specific term weighting algorithms use domain ontology to expand a certain document keywords with their synonyms to increase their weights accuracy. These weights are calculated using some statistical formulas [9].

3. Fuzzy Ontology

Ontology is "*the conceptualization of a domain into a human understandable, machine readable format consisting of entities, attributes, relationships, and axioms*". It is used as a standard knowledge representation for the semantic web [7].

Unfortunately, the conceptual formalism, supported by typical ontology, may not be sufficient to represent uncertain information commonly found in many application domains. This is due to the lack of clear-cut boundaries between concepts of the domain, i.e. a concept may be a synonym for another one with a specific matching degree. Moreover, fuzzy knowledge plays an important role in many domains that face a huge amount of imprecise and vague knowledge and information, such as text mining, multimedia information system, medical informatics, machine learning, and human natural language processing. To handle uncertainty of

information and knowledge, one possible solution is to incorporate fuzzy set theory into ontology [3], which leads to the birth of fuzzy ontology.

Although fuzzy ontology is used in many applications such as semantic web, multiagent systems, information retrieval, and many more, no standard components for fuzzy ontology are found in the literature [5]. Commonly, researchers define fuzzy ontology components according to their used application and domain. Some of such definitions are as follows:

- Fuzzy ontology is a quintuple(C, R, H, P, A), where C is a set of fuzzy concepts, R is a set of fuzzy relations between individuals, H, is a concept hierarchy, P is a set of non-taxonomic fuzzy relations between concepts, and A is a set of fuzzy axioms [6].
- Fuzzy ontology is a triple(C, I, R), where, C is a set of concepts, I is a set of individuals, and R is a set of binary relations between some elements of C and I, including two special types of fuzzy relations [8].
- Fuzzy ontology is a quadruple(C, R, P, I), where C is a set of fuzzy concepts, R is a set of binary relations, P is a set of fuzzy properties of concepts, and I is a set of individuals [10].
- Multi-Views Fuzzy Related Ontologies, MVFRO, is a couple of (f-os, f-oI), where f-os is a fuzzy ontology structure, while f-oI is the fuzzy ontology individual of concepts and relationships associated with the fuzzy ontology structure. F-os is a quintuple (C, C_R, P, T, A). C is a set of fuzzy concepts. C_R is a fuzzy relation between concepts. It can have more than one value each represents a certain view. P is a set of concept properties. T is a set of terms that express the concept c. A is a set of axioms [13].

4. Related Work

Two annotation techniques based on crisp ontology are proposed in [9]. Each technique can annotate a set of documents with a string of weighted keywords in two steps. The first step is to annotate documents with a string of keywords. This string is then entered the second step to be weighted. These weighted keywords are then stored in a relational database such that each tuple in it indicates that a document d_i is indexed by a term t_k with a weight w_j. The first annotation technique uses an NLP annotation algorithm to annotate a certain document with a string of keywords. Then, these keywords are weighted using an adapted TF-IDF algorithm. This adapted algorithm is the frequency of the occurrence of each semantic entity in the ontology or any of its associate keywords within a document. Such an algorithm takes pronoun into account. The second technique uses a contextual semantic information based algorithm to annotate a certain document with a string of keywords. Then, theses keywords are weighted using a fusion weighting algorithm.

An annotation system performing a clustering process based on a concept weight supported by crisp domain ontology is proposed in [11]. The system is divided into three major modules; document preprocessing, calculating a concept weight based on ontology, and clustering documents with the concept based. The weighting module is calculated by the following equation [11]:

$$W = Len \times Frequency \times Correlation\ Coefficient + Probability\ of\ concept \quad (4)$$

Where, W is the weight of a certain keyword, len is the length of that keyword, Frequency is times which the words appear, and if the concept is in the ontology, then correlation coefficient =1, else correlation coefficient=0. Probability is based on the probability of the concept in the document. The probability is estimated by following equation [11]:

$$P(\text{concept}) = \frac{\text{Number of Occurences of the Concept}}{\text{Number of Occurences of All Concept in Document}} \quad (5)$$

A new weighting method based on statistical estimation of a word importance for a particular categorization problem is proposed in [12]. This weighting also has the benefit that it makes feature selection implicit since useless features for the categorization problem considered get a very small weight.

5. The Proposed Term Weighting

The proposed algorithm.is a semantic based term weighting algorithm. It calculates a weight for each annotated keyword in a certain paper according to a specific view. This weight reflects to what extent such keyword represents its paper according to the specified view. In a preprocessing phase we apply a stemmer algorithm and wratelic tool to each paper. The stemmer algorithm is used to extract the stem of each word, while wratelic tool converts each pronoun to its noun. Consequently, each paper will be more convenient for applying our algorithm. Commonly, the algorithm aims to enhance the resulted weights accuracy through:

- Using a multi-view fuzzy ontology for expanding each annotated keyword in the keyword zone respecting the required view,
- Arranging the paper expanded keyword list in a descending order according to each keyword n-grams, the number of terms in each keyword,
- Replacing each pronoun with its referred noun, instead of removing it as stop word,
- For titled sections, they are annotated with their titles not with keywords listed in the paper keyword zone,
- Keywords written with different style (Bold, Italic, Underlined) will have higher weights,

- For well written papers, the proposed algorithm can be applied on each paragraph main sentence as it reflects its main idea, instead of working on the whole paper. This will reduce the execution time.

The algorithm is illustrated as in algorithm1.

Algorithm 1: Term Weighting of research paper in a certain view

Input: research paper in a certain view, a predefined multi-view fuzzy ontology

Output: research paper annotated with a set of weighted keywords in the specified view

Steps:

1. Divide the paper into different weighted zones
2. PKS= expand all keywords in the keyword zone according to the given view using the predefined fuzzy ontology
3. Arrange all PKS in decreasing order according to each keyword n-gram
4. annotate each zone with the PKS
5. for each zone, calculate the weight of each keyword in PKS
6. **For** each section
7. annotate it with its title
8. SKS= expand this annotation using the predefined multi-view fuzzy ontology
9. arrange SKS in a descending order with respect to each keyword n-gram
10. calculate the weight of each keyword in SKS
11. **End for**
12. calculate the weight of each keyword in the keyword zone through summing its value from each zone and each section

After applying the proposed term weighting algorithm, the resulted weighted keywords are stored in a relational database. As shown in figure 1, this database has three tables; *documents*, *fuzzyDocumentClasses*, and *fuzzyDocumentTerms* tables. *Documents* table stores all information needed to access any document such as, document name, and document path. Also, it gives each one a unique identifier. The annotated weighted keywords of a certain document are stored in the other two tables; *fuzzyDocumentClasses* and *fuzzyDocumentTerms*. If these keywords are represented as concepts in the used fuzzy ontology, then this annotation is stored in *FuzzyDocumentClasses* table. This table is connected with the fuzzy ontology database through the foreign key c_id (concept_id). While, if these keywords are represented as terms in the used fuzzy ontology then, these annotation type is stored in *fuzzyDocumentTerms* table. This table is also connected with the fuzzy ontology database through the foreign key t_id (term_id).

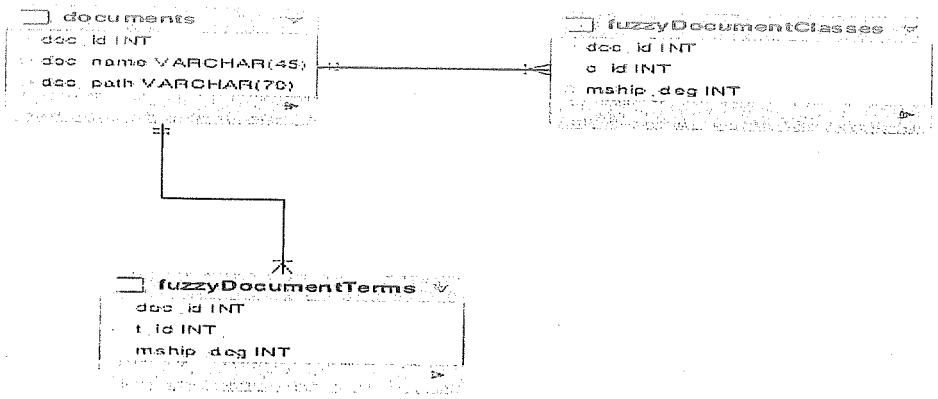


Figure1: storing documents associated with their annotated weighted keyword annotations

1. Tests and Results

The proposed Term Weighting algorithm is tested on 20 research papers and tests are compared with Fernández algorithm [9] respecting a certain view. Table 1 shows the ranked annotated keywords according to the proposed algorithm and Fernández algorithm respectively. Results show that the proposed algorithm enhances the resulted weights accuracy as the result of expanding a certain keyword using a specific domain fuzy ontology, replacing each pronoun to its related noun, arranging each annotated keyword in a descending order respecting its n-gram, annotating each section with its title, dividing the paper into different weighted zones.

Figure 2 shows the resulted weights for a sample of eight papers from those 20 ones. Each chart in this figure represents a certain paper resulted weights after applying the proposed algorithm on its paragraphs main sentence (first sentence only, the first two sentenes, the first three sentences, ...) and on its paragraphs all sentences. It is obvious from the figure that after the second line the weight of a certain keyword becomes approximately stable. Table 2 shows that working on a well written paper main sentence instead of working on all its sentences will decrease the execution time while the ranking accuracy of relative weights approximatly remains the same.

Table 1: Ranking a set of 10 research papers after applying both of Fernández and the proposed term weighting algorithms on them.

Paper Title	Fernández Keyword weights(IR)	The proposed algorithm keyword weights
An Intelligent Information Retrieval Approach Based on Two Degrees of Uncertainty Fuzzy Ontology	0.041912995	0.051360062
A Linguistic based Fuzzy Ontology Information Retrieval Model	0. 03671573	0.039556021
A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback	0.032173913	0.039420780
Using Multiple Related Ontologies in a Fuzzy Information Retrieval Model	0. 027882459	0.034245740
Semantically enhanced Information Retrieval: An ontology-based approach	0. 02617742	0.030659821
A Semantic Retrieval Framework for Engineering Domain Knowledge	0.026891671	0.021974044
A Fuzzy Ontology-Approach to improve Semantic Information Retrieval	0. 023303168	0.017925823
An Ontology-Based Information Retrieval Model for Vegetables E-Commerce	0.02200868	0.017698251
A conceptual representation of documents and queries for information retrieval systems by using light ontologies	0.020606695	0.015358944
A Survey of Concept-based Information Retrieval Tools on the Web	0.019704144	0.005698574

Table 2: Ranking a set of 10 research papers after applying the proposed term weighting algorithm on all lines and on just both of main and subordinate sentences in each paragraph

Paper title	Applying the proposed algorithm on all lines		Applying the proposed algorithm on 2 lines	
	Keyword weights	Elapsed time (SEC)	Keyword weights	Elapsed time (SEC)
Semantically enhanced Information Retrieval: An ontology-based approach	0.051360062	6 sec	0.042760313	5.6 sec
A Linguistic-based Fuzzy Ontology Information Retrieval Model	0.039556021	2 sec	0.03997887	1.5 sec
A Semantic Retrieval Framework for Engineering Domain Knowledge	0.039420780	0.8 sec	0.028512955	0.5 sec
Using Multiple Related Ontologies in a Fuzzy Information Retrieval Model	0.034245740	1.5 sec	0.024480078	1 sec
An Intelligent Information Retrieval Approach Based on Two Degrees of Uncertainty Fuzzy Ontology	0.030659821	0.5 sec	0.022557942	0.4 sec
A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback	0.021974044	2.6 sec	0.019704144	2 sec
A Fuzzy Ontology-Approach to improve Semantic Information Retrieval	0.017925823	1.7 sec	0.018235292	0.7 sec
An Ontology-Based Information Retrieval Model for Vegetables E-Commerce	0.017698251	0.7 sec	0.017818861	0.6 sec
A conceptual representation of documents and queries for information retrieval systems by using light ontologies	0.015358944	2.5 sec	0.013514333	2 sec
A Survey of Concept-based Information Retrieval Tools on the Web	0.005698574	1.2 sec	0.002023668	1 sec

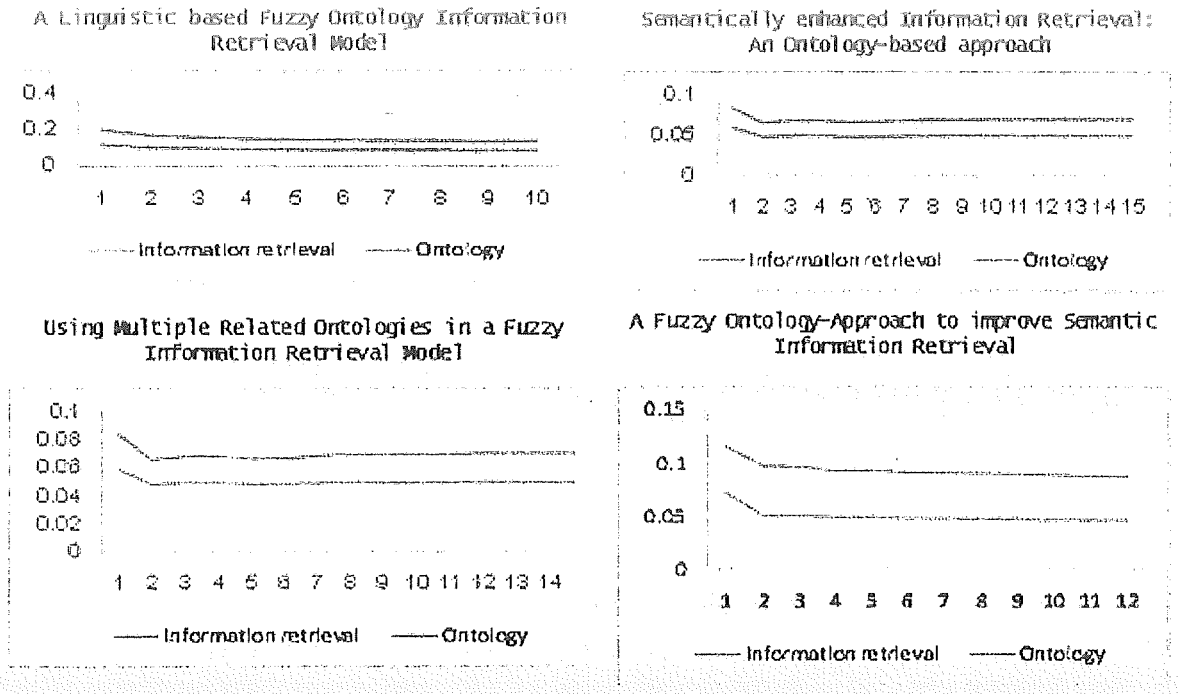


Figure 2: The relative weights of the annotated key words after applying the proposed term weighting algorithm gradually from just the main sentence up to all sentences in each paragraph.

2. Conclusion

Compared with Fernández algorithm, the proposed algorithm enhances the accuracy of weighting the annotated keywords. This is due to dividing the paper into different weighted zones, using fuzzy ontology instead of using crisp one, arranging the expanded list in a descending order respecting the number of n-grams of each keyword in it, annotating each paper section with its title, retrurning each pronoun to its refered noun. Also, applying the proposed algorithm to just main and supordinate sentences in well written papers sounds well in both of accuracy and keywords weighting time.

Commonly, in well written papers, the main sentence and its subordinate sentence in each paragraph describe and represent the paragraph main ideas. Accordingly, the weight of each keyword in a specific paragraph can be concluded from the existance of it or any of its related concepts and terms in both of the main sentence and its subordinate one. As shown in figure 2, applying the proposed algorithm to any number of lines in each paragraph starting from the main sentence preserves approximatly the relative weights of the annotaed key words. Consequently, for well written papers, applying the proposed algorithm on just both of main and subordinate sentences in each pargraph instead of working on the whole paper decreases

the execution time with approximatly the same ranking accuracy as shown in Figure 2. Assumming that, each paragraph in average contains N sentences and the average time needed for applying the proposed term weighting algorithm to a specific paragraph is T unit of time. Then, applying the proposed term weighting algorithm to just both of main and subordinte sentences of each paragraph consumes in average (2/N * T) unit of time . Accordingly, working on just both of main and subordinate sentences saves in average about ((1-2/N)*T)) unit of time and preserves, approximatly, the relative weights of the annotated keywors.

References

[1] S. Klink, K. Kise, A. Dengel, M. Junker, and S. Agne, "Document Information Retrieval," Digital Document Processing, Springer-Verlag London Limited 2007.

[2] Salton, Gerard, Buckley, and Chris, "Term weighting Approaches in Automatic Text Retrieval," Information Proceeding and Management Vol.32 (4), pp. 431-443, 1996.

[3] J. Zhai, Y. Liang, Y. Yu and J. Jiang "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce," JOURNAL OF SOFTWARE, VOL. 3, NO. 9, DECEMBER 2008.

[4] M. A. A. Leite and I. L. M. Ricarte, "Relating ontologies with a fuzzy information model," Journal Of Knowledge and Information System, pp. 619-651, 2013.

[5] F. B. Ortenga, M. D. Calvo-Flores, "Managing Vagueness in Ontologies," PHD Dissertation, Granada, October 2008.

[6] E. Sanchez, T. Yamanoi, "Fuzzy Ontologies for the Semantic Web," the 7th International Conference on Flexible Query Answering Systems (FQAS 2006), Vol. 4027, pp.691-699, 2006.

[7] Q. T. Tho, S. C. Hui, A. C. M. Fong, T. H. Cao," Automatic Fuzzy Ontology Generation for Semantic Web," IEEE transaction on knowledge and data engineering, Vol. 18, No.6, June 2006.

[8] Y. Ling, H. M. Gu, X. W, J. Q. Shi, "A Fuzzy Ontology and Its Application to Computer Games," the 4th International Conference on Fuzzy Systems and knowledge Discovery (FSKD 2007), Vol. 4, pp.442-226, 2007.

[9] M. Fernández, I. Cantador , V. López, D. Vallet, P. Castells, E. Motta , "Semantically enhanced Information Retrieval: An ontology-based approach," Web Semantics: Science, Services and Agents on the World Wide Web 9, pp. 432-452, 2011.

[10] Y. Cai, H. F. Leung, "A Formal Model of Fuzzy Ontology with Property Hierarchy and Object Membership," the 27th International Conference on Conceptual Modeling (ER 2008), Vol. 5231, pp.69-82, 2008.

[11] H. H. Tar, T. T. S. Nyunt, "Ontology-Based Concept Weighting for Text Documents," International Conference on Information Communication and Management, vol.16, 2011.

[12] P. Soucy, G. W. Mineau, "Beyond TFIDF Weighting for Text Categorization in the Vector Space Model," The 19th International Joint Conference on Artificial Intelligence, pp. 1130-1135, 2005.

[13] Z. E. Alarab, A. M. Gadallah, H. A. Hefny, "An Enhanced Model For Linguistic-based fuzzy ontology," the 47th Annual Conference on Statistics, computer sciences and operation research, pp. 49-62, 2012.

Handling OCR-Degraded Arabic Text: A Comprehensive Survey

Mariam A. Muhammad[1] and Tarek A. Elghazaly [2]

Abstract. Research work on Arabic optical character recognition (AOCR), although lagging that of other languages, is becoming more intensive than before and commercial systems for AOCR are becoming available. This paper presents a comprehensive survey of research on AOCR. It introduces the general topic of optical character recognition (OCR), and highlights the characteristics of Arabic text. In this paper, we first underline the Arabic challenges that were the reason to develop few Arabic OCR systems. Most of the research works focused on three main issues: building better Arabic OCR systems, how to improve the information retrieval effectiveness for Arabic OCR text collections and the other research try to use models and techniques for detecting and correcting Arabic OCR. Then we make a comparison between some research according to the used techniques in their research and the result of each. We conclude by pointing comments and limitation of research work on AOCR.

Keywords: Arabic OCR, Information Retrieval(IR), Post-processing, Error Correction.

1 Introduction

Optical Character recognition (OCR) is an automated process that converts the document images into editable text. The research on the OCR (Optical Character Recognition) started in the 1930's and this research focus mainly on Latin, Chinese and Japanese languages. In 1975, the research works on Arabic character recognition started and started to increase since the 1980's[Nazif,1975].
Some efforts concentrated on the pre-processing of the incoming signal and the other on post-processing of the output text for error reduction. Preprocessing depends on signal refining and adjustment, noise removal, skew detection and correction (for OCR), environment identification and subtraction ... etc. While the Post-processing depends mainly on linguistic and morphological properties of the recognized text language, and it works with syntactic analysis and language modeling.

[1, 2]Institute of Statistical Studies and Research, Cairo University, Egypt
[1]Mariam_adel@cu.edu.eg
[2]t.elghazaly@fci-cu.edu.eg

OCR systems that are often improved by performing post-processing are attempts to evaluate whether each word, phrase or sentence in the OCR output is legal and/or probable. When an illegal word or phrase is discovered (error detection), these systems usually attempt to generate a legal alternative (error correction) [Habash, 2011].

In this paper we overview the research that workson Arabic OCR and their results, we found that the research focused on building better AOCR systems, improving IR effectiveness regardless of improving text quality and spell checking and correction for Arabic text recognition.

The outline of this paper is as follows: Arabic Optical character recognition and Arabic challenges are presented in section2. Building better OCR systems is reviewed in section3. Improving IR effectiveness for Arabic OCR text collections in section 4. In section 5 we overview approaches and techniques for spell checking and correction for AOCR. We explained our comments and limitation of previous work in section 6. Finally the conclusion of this paper is in section 7.

2 Arabic Optical Character Recognition.

Arabic Optical Character Recognition (OCR) is an important and emerging application and research area. An OCR tool could be used to avoid retyping a scanned document or to convert the text images in the scanned document to a text that can be edited. Such tool takes the scanned document as an Image and recognizes the text in the Image to make it available in text format [Al.Mohtaseb, 2011].

After years of development, optical character recognition (OCR) for Latin-character languages, such as English, has been refined greatly. However, Arabic has several challenges that make OCR more difficult [Halima, 2009] and Because of these challenges, only a few systems for Arabic OCR have been developed. Some of these challenges that make the Arabic OCR difficult include (as shown in Fig.1):

- Arabic's cursive script in which most characters are connected and their shape vary with position in the word.
- The optional use of word elongations such as (ـ) and ligatures, which are special forms of certain letter sequences.
- The presence of dots in 15 of the 28 letters to distinguish between

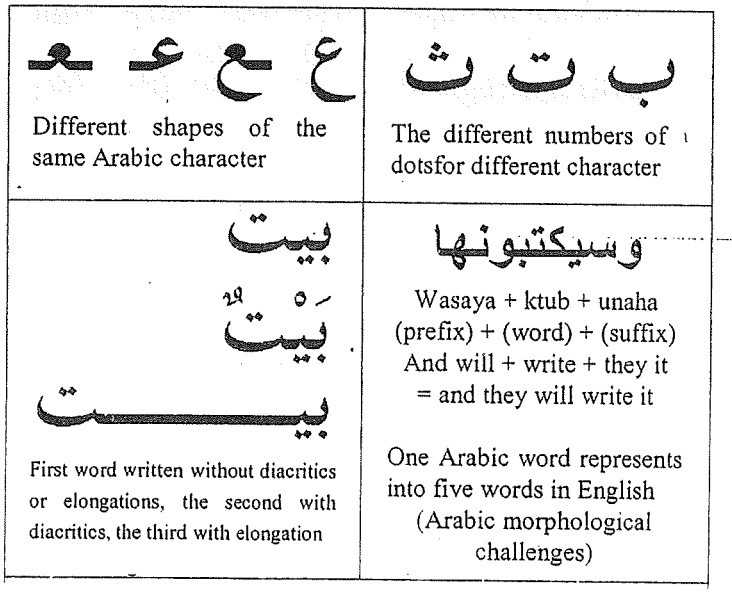


Fig.1 Arabic language orthographic and morphological challenges

different letters and the optional use of diacritic which can be confused with dirt, and speckle [Darwish, 2002].

- The morphological complexity of Arabic, results in an estimated 60 billion possible surface forms, complicates dictionary-based error correction. Arabic words mostly contain prefix and suffix and they are built from a closed set of about 10,000 root forms that typically contain 3 characters, although 4-character roots are not uncommon, and some 5-character roots do exist. Arabic stems are derived from the root forms by fitting the root letters into a small set of regular patterns [Ahmed, 2000].

Only a few systems for Arabic OCR have been developed, Because of these challenges. These systems are often improved by performing Post-Processing. And there are tries to correct OCR degraded text has been reported for many languages, including English [Harding et al., 1997], Chinese [Tseng &Orad, 2001], and Arabic [Darwish &Orad 2002].

In the next section, we will review some of the research works for building Arabic OCR systems.

3 Building Better Arabic OCR Systems

First systems for Arabic printed text were available at the market in the 1990's. Today a few systems are available at the market such as (Sakhr, IRIS, ABBYY). Compared to the high quality and wide spread usage of OCR systems for Latin characters, Arabic OCR still has to be developed. The commercial Arabic OCR systems, with Sakhr's Automatic Reader and Shonut's OmniPage being perhaps the most widely used [Kanungo et al. , 1999]. Most Arabic OCR systems use segment characters, while a few opted to recognize words without segmenting characters. A system developed by BBN avoids character segmentation by dividing lines into slender vertical frames (and frames into cells) and uses an HMM recognizer to recognize character sequences [Lu et al., 1999].Märgner and El Abedin [Märgner & El Abed, 2009] present an overview of recently published recognition systems and their accuracy until 2009.Table1 show this overview where These systems are tested with ifN/ENIT-database that use the lexicon of 937 Tunisian city/village names.

Table 1. Overview of Arabic text recognition systems (Source: [Märgner & El Abed, 2009], p.32)

Author(s)	Description	Data	Results
Amin et al. (1996)	NN	3000 characters	92%
Alimi(1997)	Nero-fuzzy	100words	89%
Dehghani et al.(2001)	multiple HMM	Farsi-Cities	71.82%
Maddouri et al.(2002)	TD-NN	70 words 2070 images	97%
khorsheed(2003)	Universal HMM	ancient documents	87%
Ama'adeed et al.(2004)	multiple HMM's	AHDB	45%
Haraty&Ghaddar(2004)	NN	2132 letters	73%
Souici-Meslati&Sellami(2004)	NN	55 words	92%
Farah et al.(2004)	ANNK-NN,fuzzy K-NN	48 words (100 writers)	96%
safabakhsh&Adibi(2005)	CD-VD-HMM	50 words	91%
pechwitz&Margner(2003) (ARAB-IfN)*	SC-1D-HMM	training: a-c, test:d	2003: 89%
		training:a-d, test: e	2005: 74.69%
Jin et al.(2005)	statistical methods	training:a-d, test: e	29.62%
Kundu et al.(2007)	VD-HMM	training:a-e, test: f	61.70
Ball (2007)	HMM	training:a-e, test: f	59.01%
Al-Hajj et al.(2006)	HMM	training:a-d, test: e	2005: 75.93%
		training:a-e, test: f	2007: 81.93%
Abdulkadr(2006)	NN (Two-Tier approach)	training: a-d, test: e	2005: 65.74%
		training: a-e, test: f	2007: 81.47%
Menasri et al.(2007)	hybrid HMM/NN	training: a-e, test: f	80.18%
Benouareth et al.(2008)	HMM	training: a-c, test: d	89.08%
Zavorin et al.(2008)	HMM	training: abce, test: d	52%
Dreuw et al.(2008)	HMM	training: a-d, test: e	80.95%
Graves &Schmidhuber (2008)	MDR-NN	training:a-e, test: f	91.43%
Kessentini et al.(2008)	HMM multi-stream	lexicon of 500 words	86.2%

The previoustable showed some research contributed in recognizing Arabic characters. Some techniques had been used in the research such as using Hidden Markov Model (HMM) in [Dehghani et al., 2001][Khorsheed, 2003][Alma'adeed et al., 2004][Safabakhsh&Adibi, 2005][pechwitz&Margner, 2003][Kundu et al., 2007][Al-Hajj, 2007][Benouareth et al., 2008][Zavorin et al., 2008][Drew eta al., 2008][Kessentini et al., 2008], Neural Network (NN) in [Amin et al., 1996][Maddouri et al., 2002][Harraty 2004][Souici, 2004][Graves, 2008], fuzzy models in[Alimi, 1997][Farah, 2004], statistical methods in [Jin et al., 2005] and combination between more than one technique in [Farah, 2004][Menasri et al., 2007]. The results show that fuzzy models have high score in recognizing the Arabic characters. Until now, much research has been proposed for obtaining on the highest quality in recognizing the Arabic characters.

4 Improving IR Effectiveness for Arabic OCR Text Collections

Many documents are available only printed. And finding printed documents or document images can be complicated process. OCR is the way to find these document images. However, OCR results often contain errors known as OCR degraded text retrieval. We reviewed the research work for improving IR effectiveness for Arabic OCR text collections in the following subsections.

4.1 OCR degraded text retrieval

Much research has been conducted to improve the retrieval effectiveness and visualization of OCR'ed documents [Darwish & Emam, 2006]. The research has been reported for many languages, including English [Harding, 1997],[Kantor, 1996],[Taghva et al., 1994],[Taghva et al., 1995]; Chinese [Tseng &Orad, 2001]; and Arabic [Darwish &Orad, 2002 a,b]. Most early studies of character Arabic text retrieval base on small test collections [Abu-salem et al., 1999],[Al-Kharashi,1994]; and others are depended on a large collection [Orad&Gey, 2002]. Several types of index terms including words, word clusters, terms obtained through morphological analysis (e.g., stems and roots), and character n-grams of various lengths. Early studies using small collections indicated that roots were the best Arabic index terms. More recent studies conducted on the larger Arabic test collection suggested that lightly stemmed words and character 3 and 4-grams result in better retrieval effectiveness than roots [Darwish &Orad $2002_{a,b}$],[Fraser et al, 2002],[larkey et al., 2002],[Mayfield et al 2001]. (Darwish &Oard $2002_{a,b}$) and (Darwish et al., 2001)reported that character 3-gram and 4-grams were the best index terms for searching OCR degraded text. (Magdy & Darwish 2008) discussed Effect of OCR error correction on Arabic retrieval. They reported that using a large language model for correcting Arabic OCR errors can minimize the need for morphological analysis. (Magdy et al., 2007) examined the effect of text fusion on improvement information retrieval effectiveness. The results show that applying fusion in conjunction with character level and language modeling based correction can further eliminate much of the

errors in the OCR text. (Elghazaly & Fahmy 2009) proposed a simulation model for Improving IR for Arabic OCR Degraded Text that depends on n-grams character replacement algorithm.
Performing Information Retrieval on OCR documents in non-English languages called Cross-Language Retrieval (CLIR) that provides some interesting and unique challenges. For handling OCR-Degraded text in CLIR, (Darwish, 2003) explained the different methods for query term replacement and he proved an approach of producing possible replacements for query terms that could have been generated by OCR proved to be a useful technique for improving retrieval of OCR-degraded text.

According to results of the research on improvement the retrieval of OCR document, we concluded that:

1. Indexing term using N-gram is better than using the roots
2. Using n-gram can minimize using morphological analysis
3. Using fusion of more than one OCR systems output can eliminate the OCR errors and then improvement the information retrieval effectiveness.

5 Spell Checking and Correction for Arabic Text Recognition

The problem of automatic spell checking and correction has been studied fordecades [Kukich, 1992]. A number of researchers published papers on spell checking and correction [Kukich, 1992] [Bowden &Kiraz 1995] [Liang 2008].
Most of the published works focused on three main issues: (1) non-word error detection;(2) isolated-word error correction; and (3) context-dependent word correction [Kukich,1992].

5.1 Non-word error detection

A non-word error results from nonexistent word in the language. The techniques that have been explored for non-word error detection can be divided into two main categories: dictionary lookup techniques and n-gram analysis (Kukich 1992).
(Zamora et al. 1981) presented a study to evaluate the effectiveness of tri-gram frequency statistics for detecting spelling errors, verifying correctly-spelled words, locating the error position within a misspelling, and distinguishing between the basic kinds of spelling errors. Their study was applied to 50,000 word/misspelling pairs collected from seven chemical abstract service databases. The tri-gram frequency table was compiled from a large corpus of text. They found that the tri-gram analysis technique was able to determine the error location within a misspelled word accurately within one character 94% of the time.

After a non-word has been detected, one or more words need to be selected as candidate corrections. So, the most common method used to correct non-word errors is the isolated-word error correction.

5.2 Isolated-word error correction

Isolated-word error correction refers to spell correction without taking any context or linguistic information in which the misspelling occurs. Work on isolated-word errorcorrection started as early as the 1960s into the present. During that time, many differenttechniques have been devised, such as minimum edit distance [Brill & Moore 2000][Magdy & Darwish 2006] [Magdy & Darwish 2010], similarity key techniques [Zobel et al. 1996], n-gram based techniques [Riseman& Hanson 1974] [Islam &Inkpen 2009a] [Islam &Inkpen 2009b], probabilistic techniques [Kemighan et al. 1990] [Church & Gale 1991], rule-based techniques [Yannakoudakis&Fawthro 1983] [Shaalan et al. 2003] and phonetic similarity techniques [Schaback& F. Li 2007]. These techniques are used to correct non-word errors.

(Kemighan et al. 1990) and (Church & Gale 1991) devised an algorithm, called CORRECT, to propose a list of candidate corrections for English language. Their approach was based on the noisy channel model. They used minimum edit distance technique to generate a set of candidate corrections that differ from the misspelled word by a single insertion, deletion, substitution or transposition.Then, a Bayesian formula was used to rank the candidate suggestions. Their proposedmethod achieved 87% correction rate. Their work is limited to the correction of wordswith a single typographical error.

(Naseem 2004) proposed a hybrid approach for Urdu spelling errorcorrection. Her approach was based on single edit distance technique to correcttypographic errors and Soundexto correct phonetics errors. She tested her approach on a test data of 744 errors, which included 444 space related errors and 280 non-space errors using dictionary of around 112,481 words. An overall accuracy score of 96.68% was reported.

5.3 Context-dependent word correction

Context-dependent word correction refers to spell correction that would correcterrors involving textual and linguistic context (Liang 2008).

Fossati and Eugenio (Fossati& Eugenio 2007) proposed a method of mixed trigramsmodel that combines the word tri-grams model and POS-tri-gram model. Theydefined confusion sets for all words in the vocabulary using minimum edit distance lessor equal to 2. One of the limitations of their approach is the lack of using a goodsmoothing technique for assigning probabilities of unseen tri-grams. Another limitationis using small training data. In addition, it skips words with less than three characters.

Haddad and Yaseen (Haddad &Yaseen 2007) presented a hybrid model for spell checking and correcting of Arabic words based on semi-isolated word recognition and correction techniques. They considered the morphological characteristics of Arabic scripts in the context of morpho-syntactical, morpho-graphemic and phonetic n-gram binary rules.

There are some hybrid efforts for integrating n-gram and morpho-syntactical analysis in spell checking (Bowden &Kiraz 1995). Although there are some efforts to describe the problems and issues involved in context-dependent spell checking for Arabic, however it is still at an early stage of development.

Magdy and Darwish (Magdy & Darwish 2010) proposed a technique for correction of OCR degraded text that is independent of character-level OCR errors. Their proposed approach did not require the training of a character error model. Instead, they used Levenshtein edit distance with uniform probability distribution for different edit operations. They used a dictionary to check the OCR'ed word and then generated the candidate corrections that are similar to the OCR'ed word. Then, they ranked all candidate corrections by using the following similarity function SED:

$$S_{ED}(w_i) = \underbrace{e^{-C.ED\,(wocr.\,wi)}}_{\text{Character model}} \cdot \underbrace{p\,(w_i)}_{\text{Unigram Model}}$$

WhereED is edit distance between OCR word (WOCR) and dictionary word (w_i), while $P(w_i)$ is the uni-gram probability of w_i in the dictionary, and C is relative to the effect of edit distance. The edit distance and tri-gram language model to select the best 5 and 10 candidate corrections are used in their work. The different forms of alef (hamza, alefmaad, alef with hamza on top, hamza on wa, alef with hamza on the bottom, and hamza on ya) to alef, and ya and alefmaqsoura to ya are normalized. Their proposed technique yielded lower correction effectiveness and required the training of a good language model matching type and style.

Shaalan et al. (Shaalan et al. 2010) proposed an approach for detecting and correcting spelling errors made by non-native Arabic learners. They used Buckwalter's Arabic morphological analyzer to detect spelling errors. Then, the edit distance techniques in conjunction with rule-based transformation approach to correct the misspelled word. They applied edit distance algorithm to generate all possible corrections. Transformation rules to convert a misspelled Arabic word into a possible word correction were applied. Their rules were based on common spelling errors made by Arabic learners. Then, they applied a multiple filtering mechanism (Morphological Analyzer Filter and Gloss Filter) to reduce the generated correction word list. They evaluated their approach using a set of test data composed of 190 misspelled words. Finally, they calculated precision and recall rates for both spelling error detection and correction to measure the performance. An 80+% recall and a 90+% precision were reported. Their test

data was designed only to cover common errors made by non-nativeArabic learners, such as Phonetic errors, Tanween errors, and Shadda errors.

More recently, (Bassil&Alwani, 2012) proposed a new post-processing technique for OCR error detection and correction based on Google's online spelling suggestion. Since Google is a giant warehouse of indexed real-world pages, articles, blogs, forums, and other sources of text, it can suggest common spellings for words not found in standard dictionaries. The proposed algorithm exploited Google's "did you mean" technology with the purpose of using query spelling suggestions to correct non-word and real-word errors in OCR output text. OCR was performed on two low-quality image documents, each in a different language: English and Arabic. Experiments undertaken showed a sharp improvement in OCR error correction rate as higher number of misspellings and linguistic errors were detected and corrected using the proposed method compared with other traditional existing ones.

Table.2 listed some of available research for correcting errors in the Arabic OCR-text, and shows the techniques they used and what their results were.

Table 2.list of some recent research for OCR correction

Author/s	techniques or algorithms					Data set	Results (Experiments Evaluation)
	Dictionary lookup	n-gram (LM)	morphology	char. segment	others		
Darwish, & Oard (2002_b)	Absent X	Present √	Present √	Absent X	---	Zad AL-Me'ad book with 2730 doc. [207 w/doc.]	Character 3&4 gram were the best index terms for searching OCR degraded text.
Magdy & Darwish (2006)	Absent X	Present √	Present √ (shallow)	Present √	---	2000 words from "the provisions of the return" book by Ibn Al-Qayim.	character segment outperform to single character and LM boosts correction, while shallow morphology had a small adverse effect
Magdy, Darwish & Rashwan, (2007)	Absent X	Present √	Absent X	Absent X	fusion of 2 & 3 different degraded versions	Eight pages, of the book (Zad Al-Me'ad). that contain 4,236 words	Applying fusion + char. level + LM can further eliminate much of the errors in the OCR text.
Magdy. & Darwish (2008)	Absent X	Present √	Absent X	Present √	---	2document collections 1. 2730 doc [207 w/doc.] of Zad AL-Me'ad 2. [383,872 articles of AFP Arabic newswire	Using a large LM can minimize the need for morph. Analysis and the indexing using char. n-grams can benefit from good correction that performs no morph. analysis.

Bassil & Alwani (2012)	Absent X	Present √	Absent X	Absent X	"did you mean" feature of Google	English Document =126 W Arabic Document = 64 W	Using Google's online spelling suggestion can improve error detection and correction.

6 Comments and Limitations of Research Work on AOCR

According to overviewing the previous work, we observed that some points are absent in almost research works such as:

- Taking into account studying more than one error in the one word. Where more research assumed founding one error.
- Taking into account correcting words with different roots that begin from two letters and to the longest number of letters.
- The working on correcting the non-word error (it isn't defined in language) in addition to real-word error (it is incorrect in the context).

And we concluded that:

- Indexing term using N-gram is better than using the roots
- Character 3&4 gram were the best index terms for searching OCR degraded text.
- Using the large n-gram can minimize the need for morphological analysis
- Combing the edit distance with word n-grams performs well on the Arabic text recognition data.
- Using fusion of more than one OCR systems output can eliminate the OCR errors and then improvement the information retrieval effectiveness.
- Using a large data set for testing or evaluation is better than using a small data set.

7 Conclusion

After years of development, optical character recognition (OCR) for Latin-character languages, such as English, has been refined greatly. Arabic, however, possesses several challenges that make OCR more difficult. In This paper, we review these challenges and survey some of the recent research aiming to improve the Arabic Optical character recognition and that focus on building better OCR systems, improving Information Retrieval effectiveness, and improving text quality through post-processing error correction. We found that most of the research works focused on three main issues(1) Building better Arabic OCR systems by using some techniques and algorithms such as: HMM, Neural Network, Fuzzy models and statistical models. The results showed that fuzzy models are the best technique for recognizing Arabic characters, (2) How to improve the information retrieval effectiveness for Arabic OCR text collections. The results showed that Character 3&4 gram were the best index terms for searching OCR degraded text. And (3) Using models and techniques for detecting and correcting Arabic OCR.Further, we make a comparison between some research according to the used techniques in their research and the result of each. We conclude by pointing comments and limitation of research work on Arabic optical character recognition (AOCR).

Acknowledgments

The authors would like to sincerely thank Dr. Mervat Gheith,Dr. Ahmed Gadallahand Dr. HeshamHefnyfor their valuablecomments.

References

Abdulkadr, A. (2006). Two-tier approach for Arabic Offline handwriting recognition.In 10th International Workshop on Frontiers in Handwriting Recognition (IWFHR), pp. 161--166.

Abu-Salem, H., Al-Omari, M., & Evens, M. (1999).Stemming methodologies over individual query words for Arabic information retrieval. JASIS, 50(6), 524–529.

Ahmed, M. ' A Large-Scale Computational Processor of Arabic Morphology and Applications'. MSc. Thesis, in Faculty of Engineering Cairo University: Cairo, Egypt, 2000.

Al-Hajj, R. (2007). , Reconnaissance hors ligne de mots manuscritscursifs per l'utilisation de systémeshybrides et de techniques

d'apprentissageautomatique,.Ph. D. thesis, l'EcoleNationaleSupérieure des Télécommunications (ENST).

Alimi, A. M. An evolutionary neuro-fuzzy approach to recognize on-line Arabic handwriting.In 4th International Conference on Document Analysis and Recognition, pp. 382--386. (1997).

Al-Kharashi, I., & Evens, M. (1994).Comparing words, stems, and roots as index terms in an Arabic information retrieval system. JASIS, 45(8), 548–560.

Alma'adeed, S., C. Higgins, and D. Elliman O_-line recognition of handwritten Arabic words using multiple hidden Markov models. In the Twenty-third SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Volume 17, pp. 75--79. (2004).

Al-Muhtaseb, H & Qahwaji, R. . 'Arabic Optical Character Recognition : Recent Trends and Future Directions'. In Ch.19 Applied Signal and Image Processing: Multidisciplinary Advancements. pp. 324-326. 2011.

Amin, A., H. Al-Sadoun, and S. Fischer.Hand-printed Arabic character recognition system using an artificial network.Pattern Recognition 29 (4), 663--675. (1996)

Ball, G. R. (2007). Arabic Handwriting Recognition using Machine Lerning Approaches. Ph.D. thesis, The Faculty of Graduate School of State University of New York at Buffalo.

Bassil. B, Alwani.M ,(2012)" OCR Post-Processing Error Correction Algorithm Using Google's Online Spelling Suggestion" . Journal of Emerging Trends in Computing and Information Sciences(cis), vol. 3, no. 1, pp. 91-99 January 2012

Benouareth, A., A. Ennaji, and M. Sellami (2008).Arabic handwritten word recognition using HMMs with explicit state duration.Journal on Advances in Signal Processing.

Bowden, T. &Kiraz, G.A., 1995.A Morphographemic Model for Error Correction in Nonconcatenative Strings.Proceedings of the 33rd annual meeting on Association for Computational Linguistics, p.7.

Brill, E. & Moore, R.C., 2000.An improved error model for noisy channel spelling correction.Proceedings of the 38th Annual Meeting on Association for Computational Linguistics ACL 00, pp.286-293.

Church, K. & Gale, W., 1991.Probability Scoring for Spelling Correction.Statistics and Computing, 1(2), pp.93-103.

Darwish K. "Probabilistic Methods for Searching OCR-Degraded Arabic Text," A PhD Dissertation, University of Maryland, College Park, (2003).

Darwish, K. &Oard, D. (2002a).' CLIR Experiments at Maryland for TREC 2002: Evidence Combination for Arabic-English Retrieval'. In TREC-2002, Gaithersburg, MD ,

Darwish, K., & Emam, O. (2006).Retrieving Arabic Printed Document: a Survey.IBM Technology Development Center, Cairo, Egypt.

Darwish,K., &Oard, D. (2002b). Term Selection for Searching Printed Arabic.Proceedings of the Twenty-Fifth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval.

Darwish,K., Doermann, D., Jones, R., Oard, D.andRautiainen M.(2001) . TREC-10 Experiments at Maryland: CLIR and Video. TREC-2001, 2001.

Dehghani, A., F. Shabani, and P. Nava Off-line recognition of isolated Persian handwritten characters using multiple hidden Markov models. In International Conference on Information Technology: Coding and Computing (ITCC). (2001).

Dreuw, P., S. Jonas, and H. Ney (2008). White-space models for offline Arabic handwriting recognition. In 19th International Conference on Pattern Recognition (ICPR).

Elghazaly, T .A & Fahmy, A. A , ' English / Arabic Cross Language Information Retrieval (CLIR) for Arabic OCR-Degraded Text ' .Proceeding the 11th IBIMA conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, pp. 208-218, LNCS 5449. 2009.

Farah, N., L. Souici, L. Farah, and M. Sellami (2004). Arabic words recognition with classifiers combination: An application to literal amounts. In Artificial Intelligence: Methodology, Systems, and Applications, pp. 420--429.

Fossati, D. & Eugenio, B., 2007.A Mixed Trigrams Approach for Context Sensitive Spell Checking.Gelbukh, A. (Ed.): CICLing 2007. LNCS, vol. 4394, pp. 623–633. Springer, Heidelberg.

Fraser, A., Xu, J., &Weischedel, R. (2002).TREC 2002 cross-lingual retrieval at BBN.In TREC-2002. Gaithersburg, MD.

Graves, A. and J. Schmidhuber (2008).Offline handwriting recognition with multidimensional recurrent neural networks. In 22sd Conference on Neural Information Processing Systems (NIPS).

Habash, N & Roth, R. M .' Using Deep Morphology to Improve Automatic Error Detection in Arabic Handwriting Recognition'. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics , pp. 875-884, 2011.

Haddad, B. &Yaseen, M., 2007. Detection and Correction of Non-Words in Arabic: A Hybrid Approach. International Journal of Computer Processing of Oriental Languages (IJCPOL), 20(4), pp.237-257.

Halima. M &Alimi.A , ' A multi-agent system for recognizing printed Arabic words '. Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt. 2009.

Haraty, R. and C. Ghaddar (2004).Arabic text recognition.International Arab Journal of Information Technology (IAJIT) 1, 156--163. [Ho et al.(1994)] Ho, T. K., J. Hull, and S. Srihari (1994). Decision combination in multiple classifier systems.IEEE Transactions on Pattern Analysis and Machine Intelligence 16 (1), 66--75.

Harding, S., W. Croft, and C. Weir. 'Probabilistic Retrieval of OCR-degraded Text Using N-Grams'. In European Conference on Digital Libraries , 1997.

Islam, A. &Inkpen, D., 2009a.Real-word spelling correction using Google Web IT 3-grams.In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 3(August), pp.1241-1249.

Islam, A. &Inkpen, D., 2009b. Real-word spelling correction using Google web 1Tngram data set. Proceeding of the 18th ACM conference on Information and knowledge management - CIKM '09.

Jin, J., H. Wang, X. Ding, and L. Peng (2005).Printed Arabic document recognition system. In Proc. SPIE-IS&T Electronic Imaging, Volume 5676, pp. 48--55.

Kantor, P., Voorhees, E. (1996). Report on the TREC-5 confusion track. TREC-5 (p. 65).

Kanungo, T., Marton, G., & Bulbul, O. 'OmniPage vs. Sakhr: Paired model evaluation of two Arabic OCR products'. SPIE Conference on Document Recognition and Retrieval (VI) (Vol. 3651,pp.109–120). (1999).

Kemighan, M., Church, K. & Gale, W., 1990.A Spelling Correction Program Based on a Noisy Channel Model.Proceedings of the 13th conference on Computational linguistics, 2, pp.205-210.

Kessentini, Y., T. Paquet, and A. M. Benhamadou (2008). Multi-script handwriting recognition with n-streams low level features. In Proc. 19th International Conference on Pattern Recognition (ICPR), pp. 1--4.

Khorsheed, M. S.(2003). Recognising handwritten Arabic manuscripts using a single hidden Markov model. Pattern Recognition Letters 24 (14), 2235--2242.

Kukich, K., 1992. Techniques for Automatically Correcting Words in Text.ACM Computing Surveys, 24(4), pp.377-439.

Kundu, A., T. Hines, J. Phillips, B. Huyck, and L. Van Guilder (2007).Arabic handwriting recognition using variable duration HMM. In 9th International Conference on Document Analysis and Recognition (ICDAR), Volume 2, pp. 644 --648.

Larkey, L., Allen, J., Connell, M. E., Bolivar, A., & Wade, C. (2002). UMass at TREC 2002: Cross language and novelty tracks. In M. V. Ellen, & P. B. Lori (Eds.), The eleventh text retrieval conference, TREC 2002 (pp 721–732). Gaithersburg, MD: NIST Special Publication 500–251.

Liang, H., 2008. Spell checkers and correctors: A unified treatment. Master's thesis, University of Pretoria, South Africa.

Lu, Z., Bazzi, I., Kornai, A., Makhoul, J., Natarajan, P., & Schwartz, RA robust, language independent OCR system. In The 27th AIPR workshop: Advances in computer assisted recognition, SPIE (Vol. 3584).(1999).

Maddouri, S. S., A. Belaid, C. Choisy, and H. AmiriModéleperceptif neuronal á vision globale-locale pour la reconnaissance de mots manuscritsarabes. In Colloque International Francophone surl'Ecritet le Document (CIFED), pp. 11--20. (2002).

Magdy, W. & Darwish, K. (2008).Effect of OCR error correction on Arabic retrieval.Inf Retrieval 11:405–425 DOI 10.1007/s10791-008-9055-y

Magdy, W. & Darwish, K., (2006).Arabic OCR Error Correction Using Character Segment Correction, Language Modeling, and Shallow Morphology.Proceedings of the 2006 conference on empirical methods in natural language processing, Sydney, Australia, (July), pp.408-414.

Magdy, W. & Darwish, K., 2010.Omni Font OCR Error Correction with Effect on Retrieval.In Intelligent Systems Design and Applications (ISDA). pp. 415-420.

Magdy, W., Darwish, K., & Rashwan, M. (2007).Fusion of multiple corrupted transmissions and its effect on information retrieval. The seventh conference on language engineering—ESOLEC'2007 (pp. 351–358).

Märgner, V & Abed, H. ' Arabic Word and Text Recognition - Current Developments ' . In Khalid Choukri and BenteMaegaard, editors, Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, April. 2009.

Mayfield, J., McNamee, P., Costello, C., Piatko, C., & Banerjee, A. (2001). JHU/APL at TREC 2001: Experiments in filtering and in Arabic, video, and web retrieval. TREC-2001. Gaithersburg, MD (p.322).

Menasri, F., N. Vincent, E. Augustin, and M. Cheriet (2007).Shape-based alphabet for off-line Arabic handwriting recognition. In 9th International Conference on Document Analysis and Recognition (ICDAR), Volume 2, pp. 969--973.

Naseem, T. &Hussain, S., 2007. A novel approach for ranking spelling error corrections for Urdu.Language Resources and Evaluation, 41(2), pp.117-128.

Nazif, A.: 'A System for the Recognition of the Printed Arabic Characters'. M.Sc. Thesis, Faculty of Engineering, Cairo University, 1975

Oard, D., Gey, F. (2002). The TREC 2002 Arabic/English CLIR Track. In TREC-2002. Gaithersburg, MD.

Pechwitz, M. and V. Märgner (2003). HMM based approach for handwritten Arabic word recognition using the IFN/ENIT-database. In 7th International Conference on Document Analysis and Recognition (ICDAR), pp. 890--894.

Riseman, E. & Hanson, A., 1974.A Contextual Postprocessing System for Error Correction Using Binary n-Grams. IEEE Transactions on Computers, 23(5), pp.480- 493.

Safabakhsh, R. and P. Adibi (2005).Nastaaligh handwritten word recognition using a continuous-density variable-duration hmm.Arabian Journal for Science & Engineering 30 (1B), 95-- 118.

Schaback, J. & Li, F., 2007.Multi-Level Feature Extraction for Spelling Correction. In IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data, Hyderabad, India, pp.79-86.

Shaalan, K., Allam, A. &Gomah, A., 2003. Towards Automatic Spell Checking for Arabic. Proceedings of the 4th Conference on Language Engineering Egyptian Society of Language Engineering ELSE'03, pp.240-247.

Shaalan, K., Aref, R. & Fahmy, A., 2010.An Approach for Analyzing and Correcting Spelling Errors for Non-native Arabic learners. In the Proceedings of The 7th International Conference on Informatics and Systems, INFOS2010. Cairo, pp. 53-

Souici-Meslati, L. and M. Sellami (2004).A hybrid approach for Arabic literal amounts recognition.Arabian Journal for Science and Engineering (AJSE) 29 (2B), 174--194.

Taghva, K., Borasack, J., Condit, A., &Inaparthy, P. (1995). Querying Short OCR'd Documents. Technical Report 94–10, Information Science Research Institute, University of Nevada, Las Vegas.

Taghva, K., Borsack, J., & Condit, A. (1994).An expert system for automatically correcting OCR output. Proceedings IS&T/SPIE 1994 international symposium on electronic imaging science and technology (pp 270–278). San Jose, CA.

Tseng, Y. and D. Oard.'Document Image Retrieval Techniques for Chinese'. In Symposium on Document Image Understanding Technology, Columbia, MD , 2001

Yannakoudakis, E. &Fawthro, D., 1983.The Rrules of Spelling Errors.Information Processing & Management, 19(2), pp.87-99.

Zamora, E., Pollock, J. & Zamora, A., 1981.The use of trigram analysis for spelling error detection.Information Processing & Management, 17(6), pp.305-316.

Zavorin, I., E. Borovikov, E. Davis, A. Borovikov, and K. Summers (2008). Combining different classification approaches to improve line Arabic handwritten word recognition. In Proceedings of SPIE - The International Society for Optical Engineering, Volume 6815.

Zobel, J., Box, G. & Dart, P., 1996. Phonetic String Matching: Lessons from Information Retrieval. In Proceedings of SIGIR-96, the 19th Annual International ACM SIGIR Conference on Research and Development inInformation Retrieval.

Access Controlled Framework for Web-based Geographic Information System

Mennatallah H. Ibrahim[1] Nermin Hamza[2] Mohamed Nour El-Deen[3] Hesham A. Hefny[4]

Abstract

Security is one of the hardest challenges that faces web-based geographic information systems (GIS), where in web-based GIS sharing data is very critical for an effective working environment. As a result, data is exposed to various security risks as it may be displayed to or altered by illegitimate users. However, there is little attention given to web-based GIS data security.

Fortunately, several access control frameworks that are built especially for web-based GIS have been proposed. This paper provides a brief overview of the work done in the area of the web-based GIS access control frameworks and it proposes a new secured web-based GIS framework.

Keywords

GIS, Access control, Vector-based data, Authorization

1. Introduction

The proliferation of geographical information services on the web is creating great opportunities for the shared use of data. Data may contain sensitive information, so such data must not be freely disclosed to or altered by illegitimate users; moreover, the users of spatial services have different roles and expertise, so there is a need to assign different privileges for operating on data.

The confidentiality and integrity of data can be ensured by enforcing access control system. Ensuring confidentiality means, preventing improper disclosure of information to unauthorized users to see this information, while ensuring integrity means, protecting data from unauthorized modifications and thus preventing unauthorized users from inserting or modifying data [2].

The power of GIS applications comes from; it's ability to relate different types of data in a spatial context. Such related data are supplied by various data providers such as governments, private companies, academic organizations, non-profit organizations, and so forth. An important problem is issued which is, how each data provider guarantee the security of its data on web. An appropriate solution of such problem can be achieved by allowing each data provider to define its own security policy.

[1, 2, 4] Department of Computer and Information Science, Institute of Statistical Studies and Research, Cairo University
[3] Department of Computer Science, Faculty of Computers and Information, Cairo University

This paper proposes a framework for access control model based on three-tier architecture. The proposed framework provides authorization rules that specify contextual permissions. It depends on an access control model which provides a mean to specify different security policies in a unique framework.

The paper is organized as the following: The paper is divided into five sections: Section 2 presents an overview of previous web vector-based access control frameworks. Section 3 provides a comparison between the previous web vector-based access control frameworks. In section 4, we propose a framework for a secured server-side web-based GIS. The work is summarized in the last section.

2. Previous Work

Previous secured web vector-based GIS frameworks can be classified into two main categories: (a) fine grained access control frameworks and (b) coarse grained access control frameworks (as shown in Figure 1). Fine grained access control frameworks allow administrators to specify authorizations against map objects at a very fine granularity level, taking into account the various spatial representations, the object dimension or the object non spatial characteristics. For example, the administrator can authorize the user to see a given object at a specific dimension but not at higher dimensions, thus hiding detailed information about the object shape, while coarse grained access control models authorize the user against all available map objects depending only on their roles.

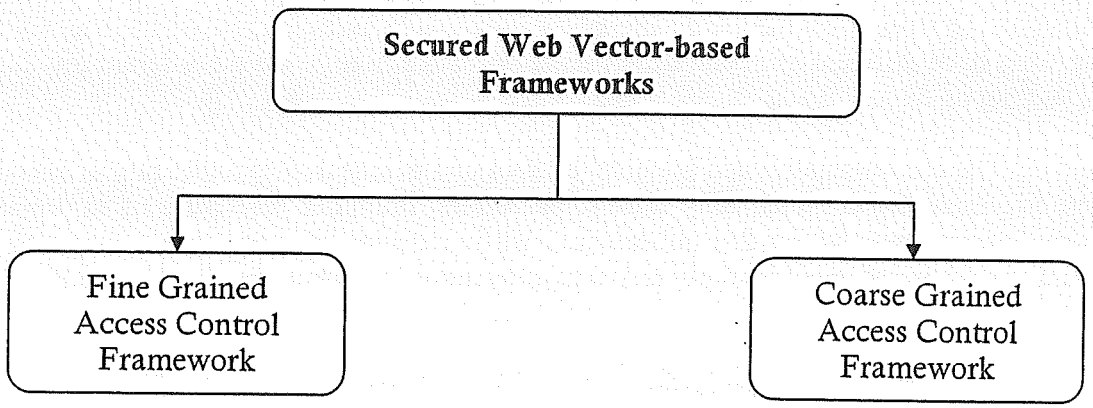


Figure 1: Secured Web Vector-based Frameworks Classification.

The first secured framework for web-based GIS was proposed in 2004 [1] by E. Bertino and M. Damiani, in which its goal is, protecting vector-based data on the web. This framework provides access control to vector data based on an extended notion of authorization rule, where its idea is to attribute a spatial dimension to the authorization rule by assigning a geographical scope that defines the spatial region in which the authorization is valid.

In Bertino framework, authorization rules are stored in a spatial database management system, where each rule is given a unique identifier whereas an attribute is defined for each of the components of the rule. Each authorization rule is defined as a tuple consisting of six fields [1]:

i. Role: it specifies the "who" of the rule;
ii. FeatureClass: it specifies the "what" of the rule;
iii. Privilege: it specifies the "how" of the rule that is the operation that is to be performed on the above Feature class;
iv. Window: it specifies the "where" of the rule and it is described through a geometric attribute;
v. Grantor: it is the role that granted the authorization;
vi. Grant Option: it is expressed as a Boolean variable; if TRUE the role is authorized to grant/revoke the rule to some other role and perform administrative operations.

Bertino framework is based on the well-known three-tier architecture which consists of presentation, application and data storage layers (as shown in Figure 2). The Application layer consists of two main services which are, the Access Control Service, and the Application Service. Access Control Service exposes and implements the operations for authorization rules checking and administration. Application Service exposes and implements the application logic and access the application data. When an operation is invoked by the user through the web browser, the Application Service interacts with the Access Control Service to checks whether the operation is authorized and can be performed or not. It has to be noticed that the application layer also includes an Authentication Service.

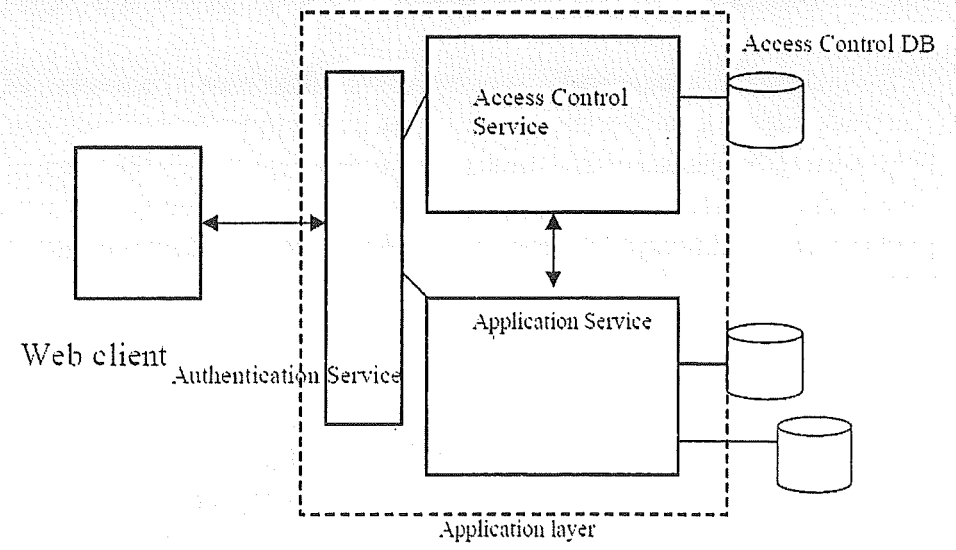


Figure 2: General architecture of the first framework [1]

The typical interaction between the web client and the services is as follows: (1) the client connects to the system through the Authentication Service. (2) The user decides which role to activate. (3) Each request from the user is then mapped onto one or more operations of the Application Service. (4) Application Service interacts with the Access Control Service to verify whether the operation can be performed and where.

The advantage of the first framework is represented in, being a first contribution to the development of secure spatial web services, so it presents the base of many other frameworks. While the disadvantage is represented in, not allowing administrator to define authorization rules for objects at a fine level of granularity.

The second framework [2] was proposed as well in 2004 by M. Damiani, D. Momini, and others. This framework is very similar to the first proposed model. It proposes an extension to the classical access control model, which is based on the notion of authorization rule. Its goal is protecting vector-based data on the web. Its central idea is assigning a geographical scope to the authorization, namely a bounded region in which the authorization is valid. The authorization rule in that framework is defined as, a tuple consisting of the same six fields as Bertino model which are, role, FeatureClass, privilege, window, grantor, and GrantOption.

Damiani framework is also organized according to the well-known three tier architecture consisting of, Presentation, Application and Data Storage layers. The Data Storage layer consists of files and database servers; the Application layer implements the operations requested by the user; and the Presentation layer is on the client side consists of either html pages or specialized programs, such as Java code, and plugs in.

Damiani framework supports two important concepts which are privileges dependencies and authorization windows constraints. In general, dependencies may exist between authorizations; this means that a privilege cannot be granted independently from some other privileges. For example, it is impossible to modify a spatial object that cannot be viewed so modifying privilege is dependent on viewing privilege. The introduction of windows in authorization rules requires revisiting the notion of dependencies between privileges in that a dependency between the privileges induces a spatial constraint between windows in the corresponding authorizations.

The advantage of the Damiani framework is represented in, being a pioneer model which provides a strong access control to spatial data especially by introducing the concepts of privilege dependencies and authorization windows constraints, in order to ensure the spatial data confidentiality and integrity. While it's disadvantages is represented in, not allowing administrator to define authorization rules for objects at a fine level of granularity.

In 2004 [3], a framework that addresses the main limitation of Bertino and Damiani frameworks was addressed by A. Belussi and E. Bertino, and others where it allows the security administrators to specify authorizations against map objects at a very fine granularity level.

In Belussi framework two new concepts are added, which are authorization sign (positive and negative authorizations) and authorization type (weak and strong authorization). Positive authorization indicates that a subject is authorized for a given privilege on a given object, whereas a negative authorization indicates that a subject is denied access to a given object under a given privilege. The authorization type specifies whether an authorization can be overridden or not. More precisely, weak authorizations can be overridden by strong authorizations.

Fine granularity is achieved in Belussi framework through both authorization type and sign. When authorization type combined with authorization sign, they create a useful mechanism for modeling exceptions. For example, if a user can update all objects of a given map except a certain object, a weak positive authorization can be granted to this user for updating all map objects and then a negative update authorization on that certain object.

In Belussi framework the authorization rule consists of eight fields instead of six which are, subject, object, privilege, window, grantor, GrantOption in addition to, authorization sign, and authorization type. The advantages of Belussi framework are, allowing one to define authorization rules for objects at a finer level of granularity and supporting the concepts of authorization type and sign which increased the framework efficiency in providing protection to the spatial data on web.

3. Previous Frameworks Comparison

Web-based GIS strongly need to be secured for several factors: (1) most of the geographical data contain sensitive information, so data cannot be freely disclosed to or altered by illegitimate users; (2) geographical information application's users have different roles and expertise, so they need to be assigned different rights for operating on data; and (3) the power of geographical information applications comes from its ability to relate different types of data in a spatial context, these related data are supplied by different data providers such as governments, private companies, academic organizations, and so forth. Each data provider need to ensure the protection of its own data when published on the Web. As a result the secured web-based GIS framework must support; strong access control model, fine grained access control and availability to allow each data provider to ensure protection of its data by its own security policy.

By comparing the three previous frameworks, it is clear that all previous frameworks provide role-based access control (RBAC) to spatial data and support the concept of authorization window. Damiani framework is better than Bertino framework where it provides more protection to spatial data through supporting to important concepts which are privileges dependencies and authorization windows constrains, while Belussi framework is better than both Bertino and Damiani frameworks, where it is more efficient in protecting the spatial data on web as it addresses the limitation of coarse granularity that was in Bertino and Damiani frameworks so more protection is provided to the spatial data through fine granularity. The differences between the three frameworks are represented in Table 1.

Table 1: Previous Frameworks Comparison

Framework	Access Control Model	Fine Granularity	Support different Security Policies
Bertino	RBAC	_	X
Damiani	RBAC	_	X
Belussi	RBAC	✓	X

4. Proposed Framework

The proposed framework is based on the well known three tier architecture which consists of presentation, application and data storage layers. The presentation layer is on the users side consists of either html pages or specialized programs, such as Java code, and plugs in. It allows users to interact with the application by sending and receiving requests through it. The application layer consists of two main services which are, the Access Control Service, and the Application Service. Access Control Service exposes and implements the operations for authorization rules checking and administration. Application Service exposes and implements the application logic and access the application data. The Application layer also includes an Authentication Service. The data Storage layer consists of files and database servers.

The proposed framework's goal is to provide security to server-side approach web vector-based GIS applications. The proposed framework is based on organization based access control model (ORBAC) [4] which provides means to specify different security policies within a unique framework, so each organization collaborated in building the application will be able to define it's own security policy and also each organization will be allowed to specify the contextual permissions.

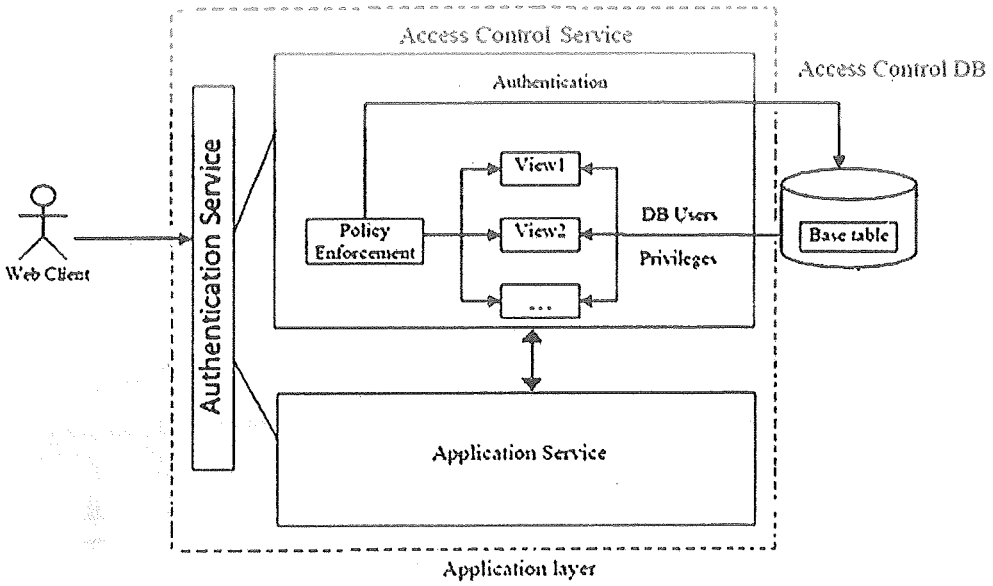


Figure 3: Proposed Framework

The proposed framework has the following components:

i. User accounts with usernames and passwords;
ii. Organizations that users belonging to;
iii. Roles that users played in the his organization;
iv. Actions that is assigned to users according to their role;
v. Views of spatial objects that have common properties;
vi. Database tables that contain spatial and non-spatial data about the spatial objects;

The views are created from the database tables in spatial database according to different access control requirements and then grant them to different roles with corresponding actions. After views satisfying access control requirements are created, the next thing is to grant access rights for users to intended views depending on the user role and organization. The proposed framework will support the fine grained access control through the created views.

The typical interaction between the user and the system will be as follows: firstly, the user will connect to the system through the Authentication Service using policy enforcement module. Secondly, an organization and a role will be assigned to the authenticated user. Thirdly, the user will be allowed to issue request(s) to the system through the presentation layer in order to perform a specific action on a specific spatial object.

By comparing the proposed model and the previous model, it is clear that the proposed model support two new important ideas that were not supported in the previous models, although they are very important to the spatial data security on web. The first idea presented in, the ability of the proposed framework to provides means to specify different security policies within a unique framework. So each data provider collaborated in building

the web-based GIS application can define its own security policy. The second idea presented in, the ability to specify contextual permissions. Furthermore, the proposed framework supports the fine-grained access control like Belussi model but through views instead of authorization type and sign.

The proposed framework advantages:

1. The framework depends on ORBAC model which provides means to specify different security policies within a unique framework.
2. The framework supports fine granularity access control to data through views.
3. The framework provides authorization rules that specify contextual permissions.

The proposed framework limitations:

1. The proposed framework is restricted to permissions, as it does not include the possibility to specify neither of prohibitions, obligations and recommendations.

5. Conclusion and Future Work

Security is very important and critical to web-based GIS. Access control is required to keep its spatial data confidentiality and integrity. Access control frameworks for web-based GIS can be classified can be classified into fine grained and coarse grained models. In this paper, we propose a framework for secured web-based server-side GIS based on organization based access control model. The goal of this framework is to ensure the security of spatial data in web-based server-side GIS.

The proposed framework will be implemented and evaluated in the near future. A case study that uses the proposed framework will be carried out to ensure the proposed framework feasibility and effectiveness.

References

[1] E. Bertino & M. L. Damiani. *"A Controlled Access to Spatial Data on Web"*. 7th AGILE Conference on Geographic Information Science, Heraklion, Greece. 2004. Pages 369-377.

[2] E.Bertino, D. Momini & M. L. Damiani. *"An Access Control System for a Web Map Management Service"*. IEEE Proceedings of the 14th International Workshop on Research Issues on Data Engineering: Web Services for E-Commerce and E-Government Applications. 2004. Pages 33-39

[3] A. Belussi, E. Bertino & B. Catania, M. L. Damiani, and A. Nucita. *"An Authorization Model for Geographical Maps"*. In Proceedings of the 12th International Symposium of ACM GIS, Washington DC, USA, November 2004, pages 82–91.

[4] A. Abou El Kalam & R. El Baid. "Organization based access control". Proceedings of the 4th IEEE International workshop on Policies for Distributed Systems and Networks. 2003. Page 120

Digital Change Detection in Burullus Protected Area Based on Temporal Multispectral Imageries

H. Farouk* and B. Abdellatif**

Abstract

Timely and accurate change detection of Earth's surface features is extremely important. This enables understanding relationships and interactions between human and natural phenomena, which in turns promotes better decision making. In recent decades, remote sensing data are primary sources extensively used for change detection. Techniques based on multi-temporal, multi-spectral, satellite-sensor acquired data have demonstrated potential as a mean to detect, identify, map and monitor ecosystem changes, irrespective of their causal agents. Many change detection techniques have been developed. However, no single approach is optimal and applicable to all cases. In practice, different algorithms are often compared to find the best change detection results for a specific application. The problem of land cover change detection can be tied to previous work in two broad themes: change detection in general (in the fields of statistics, image processing, etc.), and the specific land cover change problem studied in earth science. This paper addresses how both themes can be mixed out to answer the question of change detection spatially, quantitatively and statistically. This paper aims to utilize remote sensing (RS) and geographic information system (GIS) for mapping and detecting land cover changes. This paper does not discuss the reason for these changes. It only addresses the fact that there are some changes and finds out them.

The studying area is Burullus protected area (BPA) that contains Burullus lake (BL). BL is the second largest lake in the Egyptian Nile Delta, that makes this area, relatively, important compared to other wetland areas. This has qualified BPA to be registered as Ramsar site on 1971, and identified as an Important Bird Area (IBA) on 2001 by the Birdlife International. Currently BL faces many threats because of land reclamation for agriculture and fish farming. The paper highlights how BL land cover changes since the 1970s based on temporal multispectral Landsat series of imageries (TMLSI) containing one of the latest Landsat 8 imageries, launched in February 11, 2013. It was found that, changes in BL and BPA are a cumulative process. The results need experts of conservation and environment-related to explain why these changes happened and to develop a plan for sustainable use of the site.

Keywords

Change Detection, Image Processing, GIS, Remote Sensing, Temporal Multispectral Imageries, Landsat, Classification, Wetland, Burullus Protected Area (BPA).

1. Introduction

Landsat Multispectral data provide the longest duration archive of moderately high spatial resolution satellite imageries for monitoring the types and rates of land-surface change imposed by human activity. Although the consistent data characteristics

* *Egyptian Environmental Affairs Agency, Egypt.* HaithamFarouk@yahoo.com
** *National Authority for Remote Sensing & Space Sciences, Egypt.* Bassam.AbdelLatif@narss.sci.eg

of Landsat data enable ready production of change images, the procedures of change image production can also be implemented on multiple data sets of non-similar data characteristics, allowing combination of multiple sensors [1].

Growth of the human population, rising consumption and rapid globalization have caused widespread degradation and disruption of natural systems, especially in the freshwater realm. Freshwater ecosystems have lost a greater proportion of their species and habitat than ecosystems on land or in the oceans and they face increasing threats from dams, water withdrawals, pollution, invasive species and overharvesting. Freshwater ecosystems and the diverse communities of species found in lakes, rivers, and wetlands may be the most endangered of all [15]. The Egyptian delta ecosystem no longer receives a yearly input of sediments and nutrients from upstream. Consequently, the soils of the floodplains are poor and large amounts of fertilizers are applied to the land each year. Run-off of fertilizers and dumping of wastewater and sewage sludge lead to the accumulation of trace elements in the sediments of the delta [17].

Wetlands are of ecological importance due to their hydrologic attributes and their role as ecotones between terrestrial and aquatic ecosystems [13]. They function as downstream recipients of water and waste from both natural and human sources, and have been often transformed to dry lands for agriculture and human settlements. Moreover, for BL, Nile river control schemes have often caused the loss of wetlands. The attributes of wetlands include high productivity; sources, sinks and transformers of numerous chemical, biological and genetic materials; and valuable habitats for fisheries, wildlife and birds. Conservation associations worldwide have noted and described the alarming changes in these important habitats [10].

Wetlands are believed to play a considerable role in the global climate change by acting as a source of atmospheric greenhouse gases such as methane and a sink for carbon and nitrogen. A key aspect of wetland management includes mapping and monitoring so that public managers and decision makers can be provided with reliable and relevant information [18].

Ecosystems are in a state of permanent flux at a variety of spatial and temporal scales all around the world. Causes of these fluxes can be natural as well as anthropogenic, or may be a combination of the two. Moreover, scientific evidence clearly points to the fact that impacts of, for example, global change on land surface attributes are not uniformly distributed on the face of the Earth. The fact that sustainability has become a primary objective in the present-day ecosystem management has as one of its consequences the continuous need for accurate and up-to-date resource data. What is more, where it concerns large-area processes, any in-depth understanding of the changes has to be based on an accurate monitoring of land surface attributes over at least few decades. At such regional scales, monitoring poses a number of methodological challenges [11].

Recent developments in geographical information system (GIS) and Remote Sensing Technologies (RST) provides valuable tools to assist with monitoring, inventory and management of wetlands [21]. Change detection is the process of identifying differences in the state of an object or phenomenon by observing it at different times. Timely and accurate change detection of the Earth's surface features provides the foundation for better understanding relationships and interactions between human and natural resources to better manage and use resources [1]. Due to the importance of monitoring change of the Earth's surface features, research of change detection techniques is an active topic, and new techniques are constantly developed.

However, no single approach is optimal and applicable to all cases. In practice, different algorithms are often compared to find the best change detection results for a specific application [2].

Many remote sensor systems and environmental parameters must be considered whenever performing change detection. Failure to understand the impact of the various parameters on the change detection process can lead to inaccurate results [19]. Four major results should be provided for a good change detection research: area and rate change, spatial distribution of changed classes' types, change trajectories of land-cover types, and accuracy assessment of change detection results. Three major steps are involved when implementing a change detection technique: image preprocessing including geometrical rectification and image registration, radiometric and atmospheric correction, and topographic correction if the study area is in mountainous regions, selection of suitable techniques to implement change detection analyses, and accuracy assessment. Nine factors are involved in the accuracies of change detection: precise geometric registration between multi-temporal images, calibration or normalization between multi-temporal images, availability of quality ground truth data, the complexity of landscape and environments of the study area, change detection methods or algorithms used, classification and change detection schemes, analyst's skills and experience, knowledge and familiarity of the study area, and time and cost restrictions. Four important aspects of change detection for monitoring natural resources should be considered before implementing change detection: detecting if a change has occurred, identifying the nature of the change, measuring the area extent of the change, and assessing the spatial pattern of the change. Five categories of causes that influenced land-cover change: long-term natural changes in climate conditions, geomorphological and ecological processes such as soil erosion and vegetation succession, human-induced alterations of vegetation cover and landscapes such as deforestation and land degradation, inter-annual climate variability, and the greenhouse effect caused by human activities [13].

The problem of land cover change detection can be tied to previous work in two broad themes: change detection in general (in the fields of statistics, image processing, etc.), and the specific land cover change problem studied in Earth Science [12]. This paper addresses how both themes can be mixed out to answer the question of change detection spatially, quantitatively and statistically.

2. Site Selection and Data Sources

BPA contains BL, which is the central of five principal coastal lagoons of Northern Egypt. Manzala lake and Bardawil lake are located on the east, whereas Idku lake and Mareotis lake are located on the west of BL. BPA is a part of the governorate of Kafr El-Sheikh in the northern part of the Delta that lies between the two branches of the Nile. BL is separated from the Mediterranean Sea by a 65Km long sand bar; the middle section of the bar is narrow and is cut by an inlet that connects the sea and BL. It is the second largest lake in Egypt, which makes this area, relatively, important compared to other wetland areas in the Nile Delta.

BPA has been declared a conservation site under the Egyptian law 102/1983 in May 1998, including BL and the sandbar within the boundaries of the protected area. BL was registered as Ramsar site on 1971, and was also identified as an Important Bird Area (IBA) by the Birdlife International on 2001. Figure (1) shows the map of BPA site

location in Egypt relative to Landsat path/row. The value of BPA as a breeding area for water-birds is high, both in respect of an Egyptian standard and on an international scale. The biodiversity of BPA is relatively high (more than 700 known species). It can be considered as a biodiversity hotspot [3].

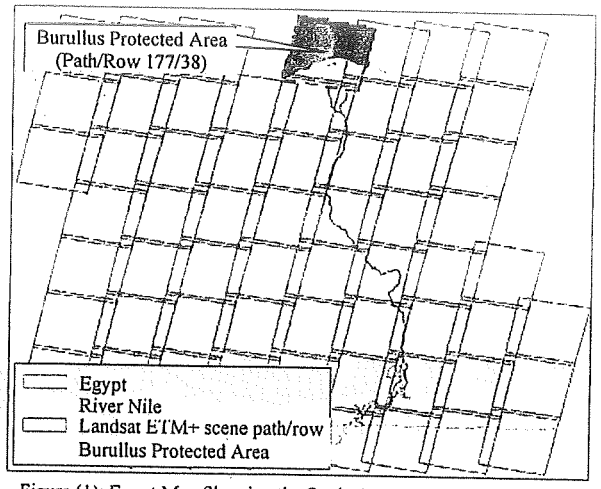


Figure (1): Egypt Map Showing the Study Areas of Burullus Protected Area.

BPA along the deltaic Mediterranean coast of Egypt is one of the largest Egyptian coastal wetlands located in the Nile Delta between the two Nile River branches: Damietta and Rosetta. Its importance arises not only from its fishing resources but also from being a habitat for domestic fauna and route for winter migratory birds [18]. Figure (2) shows a map for BPA with some photos of its habitat.

BL is separated from the Mediterranean Sea by a narrow sandy bar covered by sand dunes, sabkhas and sand flats. The only connection with the sea occurs through a narrow strait at its northeast section making the northern part of BL of much saline water than the rest of the water body [22].

In addition, there are six drains collecting agricultural and municipal wastes and pouring them into the lagoon. Water salinity at the southern part is mainly brackish (~2000 ppm) due to mixing of the fresh water from agricultural drainage with saline

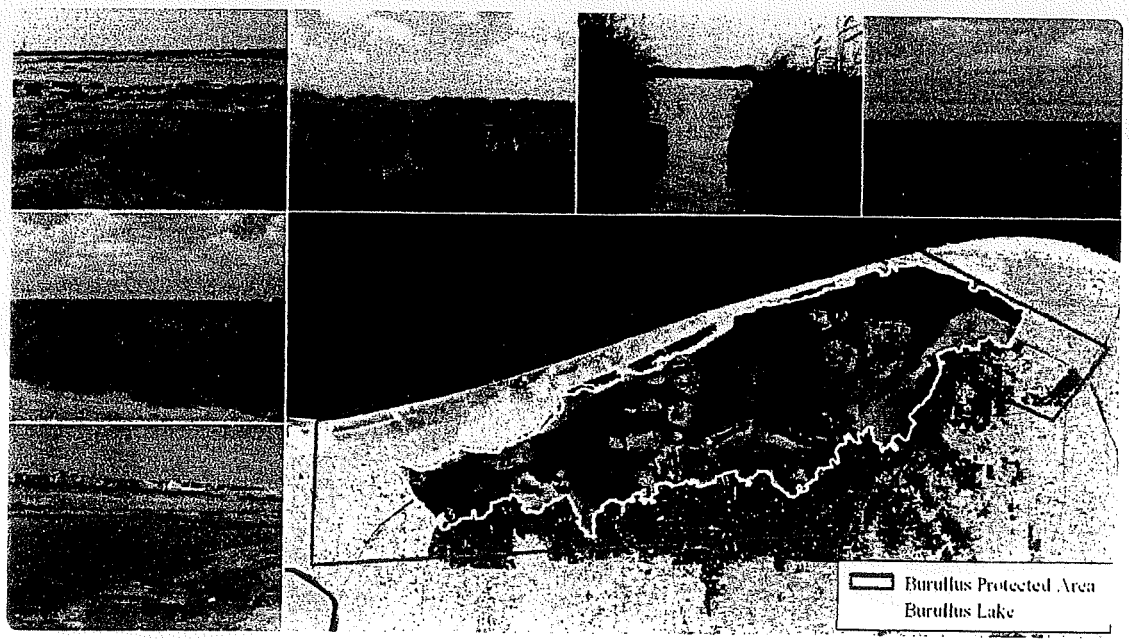


Figure (2): False-color composite of Landsat 8 OLI 2013, Bands 7,5,3 for BPA and Some Photos of BL Habitat.

water. The southern sector of BL contains reed swamp vegetation, mainly *Phragmites australis* and *Juncus* spp. This sector has witnessed drying for agricultural expansion. BL is generally shallow, where water depth ranges from 0.5 to 1.5 m for the majority of the lagoon. Climate conditions prevailing at this coastal part is generally arid, where annual precipitation does not exceed 200 mm. Northwestern winds is most frequent throughout the year [18].

A National Working Group, was established under the "Conservation of Wetland and Coastal Ecosystems in the Mediterranean Region" project (MedWetCoast Project) funded from United Nations Development Programme (UNDP) and Global Environment Facility (GEF) and implemented by Ministry of Environment, Egyptian Environment Affairs Agency. The selection of BL was due to the increasing threats on the existing natural resources; (a) A serious effect of increased flow of nutrients rich fresh waters from irrigation drains into BL is the rapid encroachment of *Phragmites* along the southern shore, which has reportedly covered 25-30% of BL area. Water hyacinth is present where the drainage canals reach BL, however this species has low salt tolerance, which prevents it colonizing larger areas, (b) Bird hunting is reported to be common, and quail nets are set along the length of coastal dunes, (c) Land reclamation for agriculture and for fish farms has reportedly affected 30% of the original BL area since 1970, (d) The inflow of drainage water has increased phosphate levels by 4 times between 1970 and 1987, and blooms of blue-green algae have been recorded at Baltim (probably linked to sewage outflows). The enrichment of the fish farms with waste, urea and superphosphates further increases fertilizer flows into BL, (e) Analyses show that the fishery is over-exploited, with a very high percentage of small fish in the catch, and (f) With the construction of the International Highway along the Mediterranean coast, there are plans for tourism development and agricultural reclamation along the sandbar of BL, which could, without adequate mitigation measures, lead to further pollution of BL and destruction of sensitive habitats for fauna and flora [3].

BPA is located at path/row 177/38 for Landsat TM, ETM+ and OLI and path/row 190/38 for Landsat MSS. Surveys of the available online Landsat imageries and clouds-free of the selected site of study for different dates; taking into consideration the conditions of the change detection, results of a TMLSI of seven different dates. That is why these dates are selected for this study. Table (1) shows the metadata of these

Serial	*Path*	*Row*	*Landsat*	*Scene ID*	*Image Date*	*Resolution (m)*
1	190	38	Landsat 1 MSS	p190r38_1m19720831.MSS-EarthSat-Orthorectified	31/08/1972	60x60
2	177	38	Landsat 5 TM	LT51770381984239XXX04	26/08/1984	30x30
3	177	38	Landsat 5 TM	LT51770381999120XXX02	30/04/1999	30x30
4	177	38	Landsat 7 ETM	LE71770382001357SGS00	23/12/2001	30x30
5	177	38	Landsat 7 ETM	LE71770382003107ASN00	17/04/2003	30x30
6	177	38	Landsat 7 ETM	L71177038_03820050305.ETM-GLS2005	05/03/2005	30x30
7	177	38	Landsat 8 OLI	LC81770382013254LGN00	11/09/2013	30x30

Table (1): Metadata of Used Landsat Scenes of the Study Area of BPA.

scenes. All these images have been downloaded from USGS using Earth Explorer and Glovis web applications. All image processing has been conducted using Intergraph ERDAS Imagine v. 2013 and ESRI ArcGIS v. 10.2 software.

3. Methodology

The next flow chart in Figure (3) summarizes the used methodology. The preprocessing is done first on the TMLSI. Then vector images are extracted by on-screen digitizing. The TMLSI is then clipped to BPA and BL. Unsupervised classification using ISODATA algorithm is the applied. Ultimately, land cover changes were found spatially, quantitatively and statistically.

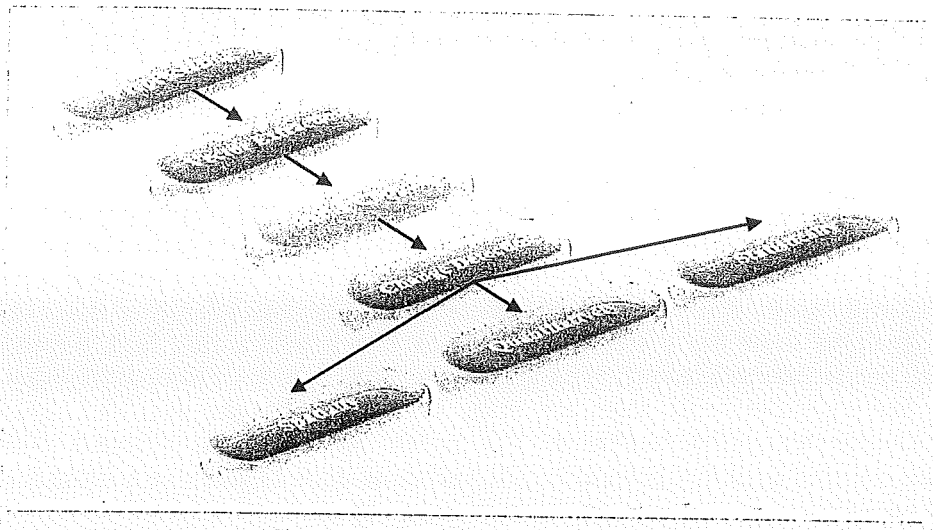


Figure (3): Flow Chart of the Used Methodology.

3.1 Preprocessing

The derivation of change information from Landsat data generally consists of co-registering the data of two or more images of the same area acquired at different times; adjusting the radiometric properties of the data to normalize for varying observation and atmospheric conditions; implementing a change detection method on the combined data sets; and producing an output product that can effectively convey land-surface change on an image or in statistical basis. Therefore, all TMLSI were originally rectified to the Universal Transverse Mercator (UTM) projection. Then, preliminarily geometric and radiometric corrections are done. Image to image registration was done for all used images. Keeping the root mean square error (RMSE) less than 0.5 pixels, an image to image transformation model (third order polynomial) and nearest neighbor resampling was calculated.

3.2 Vector Extraction

Vector extraction is to extract the boundaries of a feature in a vector image format. BL borders were extracted from each individual image of the TMLSI. Because of the disturbed shape of BL and the vegetation that lies at its shoreline, there is no technique to extract BL borders except the on-screen-digitizing and creating a shapefile vector image as shown in Figure (4a).

3.3 Raster Clipping

Raster clipping is to subset the raster image to only the studying area. BL and BPA were clipped from each individual image of the TMLSI as shown in Figure (4b) and Figure (5a), respectively. Ultimately, GIS-based operations were undertaken to calculate the area of BL for each individual image of the TMLSI. The last column of Table (2) and Table (3) show the area of BL and the area of BPA without BL, respectively. Where, the BPA is 913.92 km^2.

3.4 Classification

Image classification is an important part of the remote sensing and image analysis. In some instances, the classification itself may be the objective of the analysis. In Fact, classification of landuse from remotely sensed data produces a map-like image as the final product of the analysis. The image classification therefore forms an important tool for examination of the digital images [7]. The analyst must select a classification method that will best accomplish a specific task. At present, it is not possible to state which classifier is best for all situation as the characteristics of each image and the circumstances for each study vary so greatly. Therefore, it is essential that each analyst understands the alternative strategies for image classification so that they may be prepared to select the most appropriate classifier for the task in hand [8].

The first step involves the classification of each of the TMLSI independently, followed by a comparison of the corresponding pixel (thematic) labels to identify areas where change has occurred. There are two common approaches of multispectral classification: supervised and unsupervised classification. For the first one; an analyst selects sites of a known cover type, while for the second one; clusters of pixels are generated based on spectral similarities. Both supervised and unsupervised classification requires human expertise to generate the classified imagery. The comparison of separately classified images can be carried out visually, or by computer. Computers are better at quantitative analysis, but humans are able to discern patterns and shape much better, and the effect of miss-registrations may be much-reduced [20].

The unsupervised classification was selected for classification of the study area of BPA. It is done by ISODATA clustering algorithm. The ISODATA uses the minimum temporal distance formula to form clusters. Arbitrary cluster means are used as a signature set, and each time the clustering repeats, the means of these clusters are shifted. The new cluster means are used for the next iteration. The ISODATA algorithm repeats the clustering of the image until: either a maximum number of iterations has been performed, or a maximum percentage of unchanged pixels has been reached between two iterations. ISODATA is designated by 1) the number of classes {15}, 2) the maximum number of iterations {6}, 3) the maximum number of times a pixel can be moved from one cluster to another with each iteration {0.950}, 4) the minimum distance from the mean {calculated for each image - software wise}, and 5) the maximum standard deviation allowable {1} [3].

The resulted classes were investigated using the spectral signature profile and visual image overlay. Signatures of the same behavior are aggregated into just one class. It was decided to aggregate all these classes into just three classes. Tests, that can help to determine whether the combined signatures are true representatives of the classes of interest, were performed. Transformed divergence test; which is a measure of the signature separability, gives results greater than 1900 for all classes, which means that these classes can be well separated [9]. Field verification assists to determine the

features of the studying area. In fact, there are vegetation, agriculture, land, urban, fish farms and water surface. Therefore, these classes are; water surface of BL and fish farms, vegetation and agriculture, and urban, island and bare land; for which it was decided to interpret the graphs and figures of these classes in this study by blue, green and red colors, respectively. Absolutely, water surface, vegetation and islands located into BL. However, fish farms, agriculture urban and bare land located outside BL. Therefore, classes can fortunately be distinguishable in and out BL within BPA. That is why this study divided the imageries into two parts; BL and BPA without BL. Figure (4c) and Figure (5b) show the classification images series of the TMLSI for BL and BPA, respectively. In addition, Figure (8) and Figure (9) identify graphically the individual classes of the TMLSI and its corresponding area for BL and BPA, respectively.

3.5 Digital Change Detection Technique

A successfully implementation of change detection analysis using remotely sensed data requires careful considerations of the remote sensor system, environmental characteristics and image processing methods. The temporal, spatial, spectral and radiometric resolutions of remotely sensed data have a significant impact on the success of a remote sensing change detection project. Important environmental factors include atmospheric conditions, soil moisture conditions and phonological characteristics. Among various requirements of preprocessing for change detection, multitemporal image registration and radiometric and atmospheric corrections are the most important. The importance of accurate spatial registration of multi-temporal imagery is necessary because largely spurious results of change detection will be produced if there is misregistration [2].

In the context of spatial change detection, Figure (10a) shows an overlay map of the vector images of BL of the TMLSI. This shows how the shoreline of the BL changes over the years of the TMLSI. Quantitatively, the changes of BL area between every consecutive pair of dates of the TMLSI is shown in Table (4) and Figure (11). In details, Figure (10b) and Figure (10c) address the changes between only the first and last dates; 1972 and 2013, of the TMLSI. This is indicated by the classified images of BL dated 1972 and 2013 with the changed part of BL compared to 1972 as shown in the associated tables of the figures, respectively. However, in the context of statistics, Table (2) and Table (3) show the areas and their percentages of the BL classes and BPA without BL classes over the years of the TMLSI.

4. Discussion and Results

Changes in classes of TMLSI of BL indicate the area and trend over the last 40 years, where the blue color refers to water surface, green color to vegetation and red color to land. They also show that the majority of vegetation assemblages were located at the southern west part of BL on 1972 while during the last 29 years there was a shift in location and area of the vegetation toward the central and southern east part of BL. Moreover, there is a clear decrease in water surface from 450.24 km^2 (78.63%) on 1972 to 289.45 km^2 (63.99%) on 2013 (total amount of water surface lost 160.79 km^2 (35.7%) compared to area of 1972). Also, there is an irregular pattern in vegetation cover from 109.03 km^2 on 2003 to 209.41 km^2 on 1984 (mean area coverage over the study period 141.71 km^2) and an irregular pattern in land cover from 1 km^2 on 2005 to 6.57 km^2 on 1984 (mean area coverage over the study period 4.04 km^2).

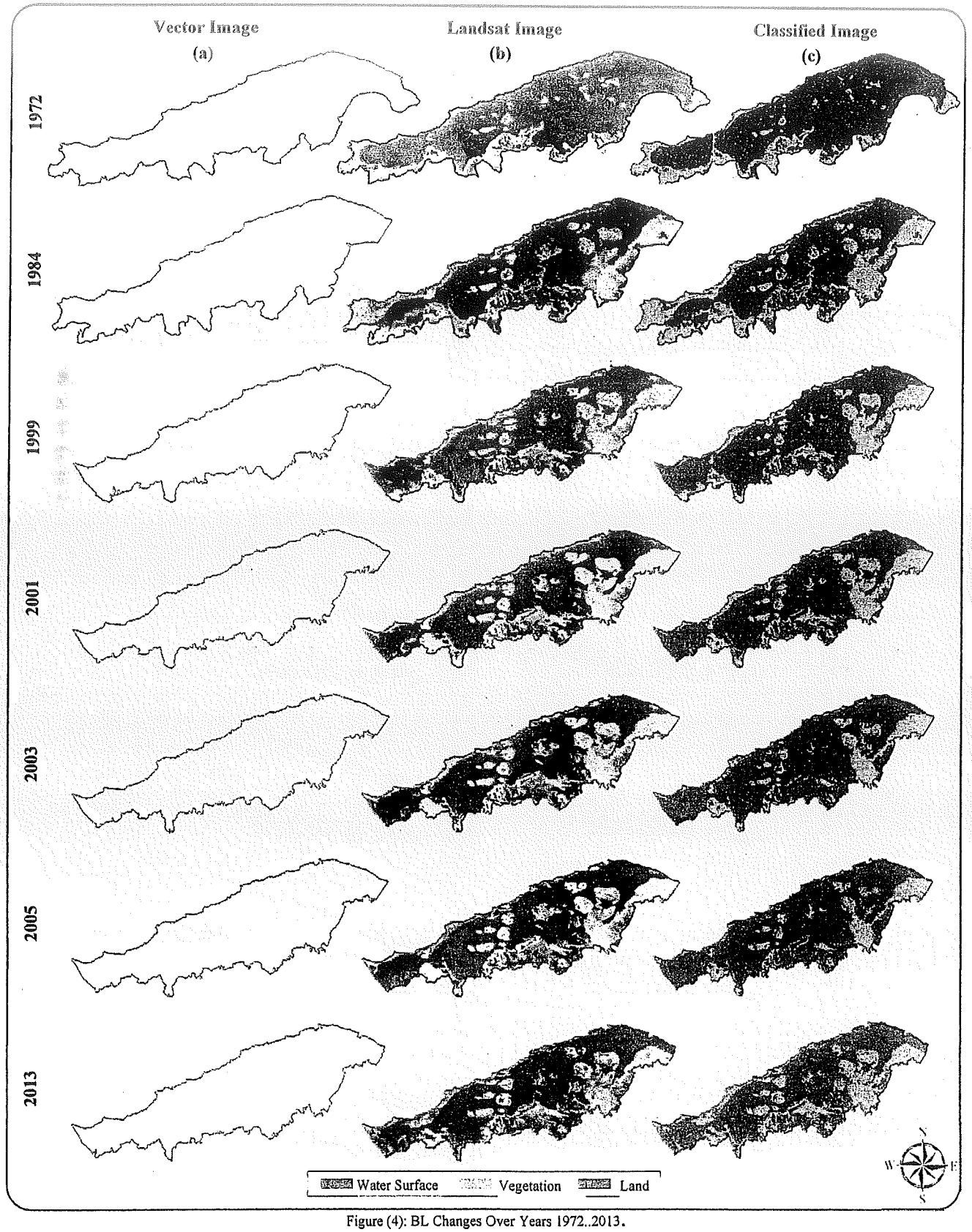


Figure (4): BL Changes Over Years 1972..2013.

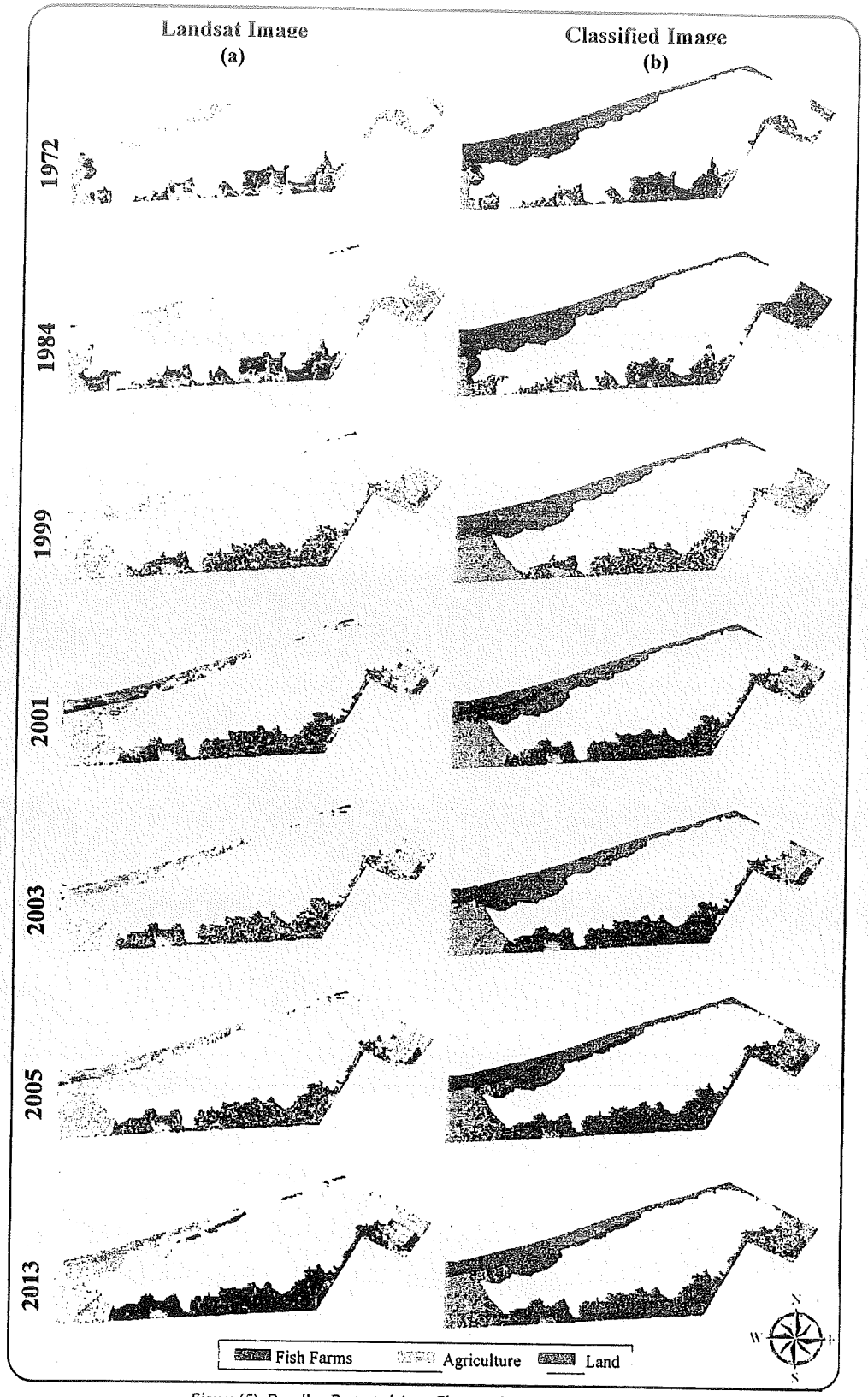


Figure (5): Burullus Protected Area Changes Over Years 1972..2013.

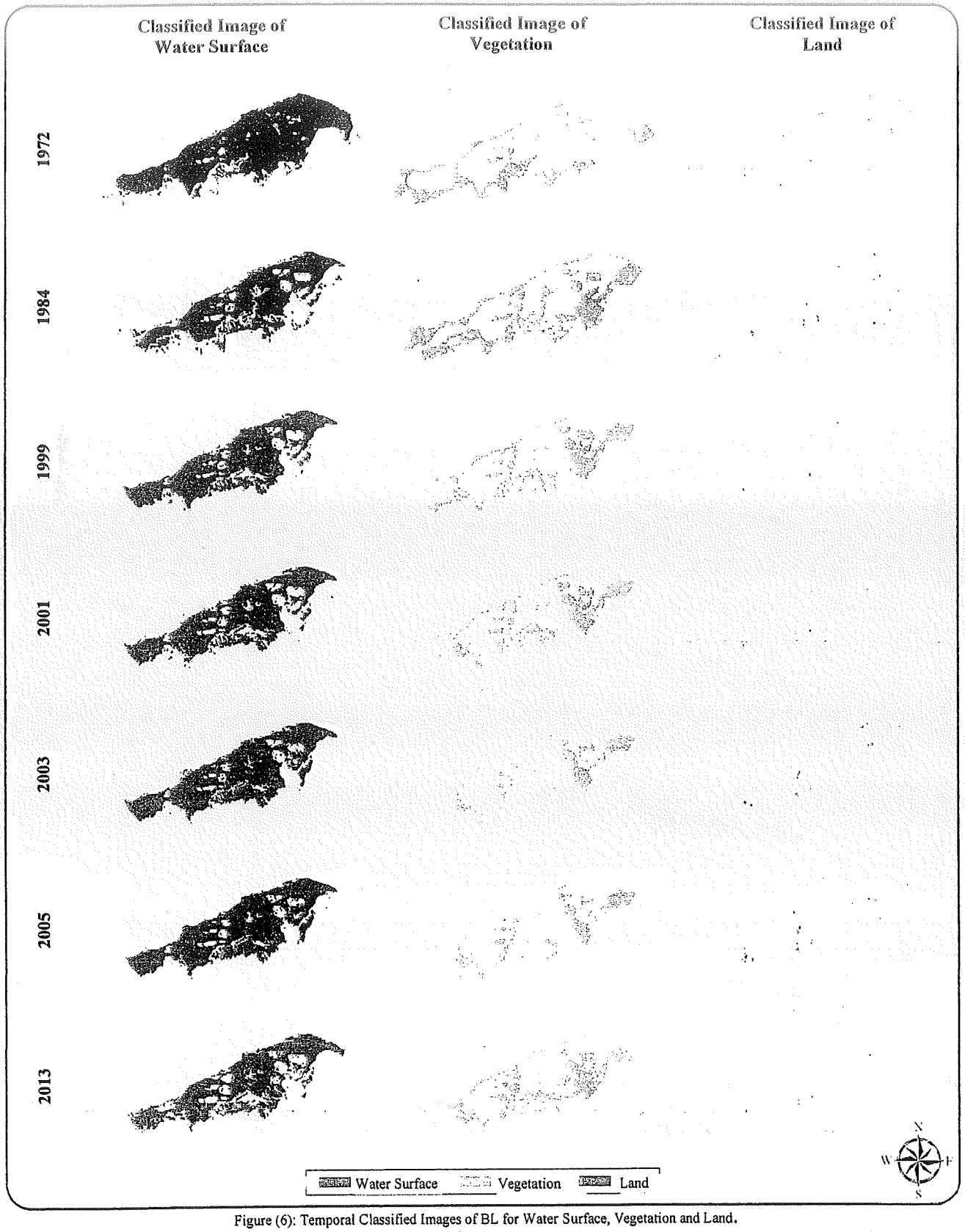


Figure (6): Temporal Classified Images of BL for Water Surface, Vegetation and Land.

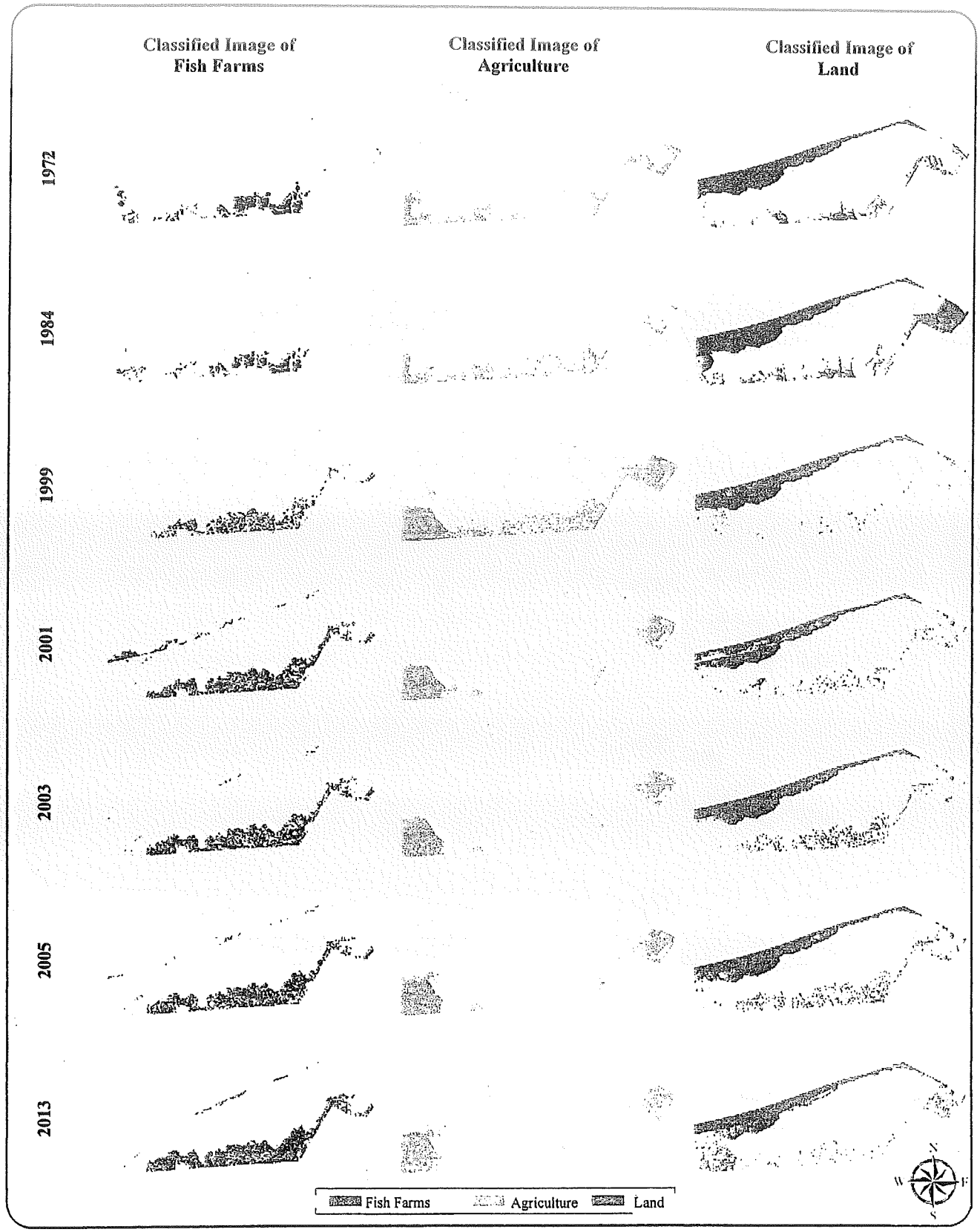


Figure (7): Temporal Classified Images of Burullus Protected Area for Fish Farms, Agriculture and Land.

Year \ Class		Water Surface Area (Km^2)	Vegetation Area (Km^2)	Land Area (Km^2)	BL Area (Km^2)
1972	Area	450.24	117.13	5.23	572.60
	%	78.63%	20.46%	0.91%	
1984	Area	314.16	209.41	6.57	530.15
	%	59.26%	39.50%	1.24%	
1999	Area	323.57	145.30	2.07	470.93
	%	68.71%	30.85%	0.44%	
2001	Area	318.45	141.61	4.27	464.33
	%	68.58%	30.50%	0.92%	
2003	Area	345.64	109.03	6.35	461.01
	%	74.97%	23.65%	1.38%	
2005	Area	348.14	109.44	1.00	458.57
	%	75.92%	23.87%	0.22%	
2013	Area	289.45	160.08	2.79	452.32
	%	63.99%	35.39%	0.62%	

Table (2): Areas of the Resulted Classes of BL Using Unsupervised Classification.

Year \ Class		Fish Farms Area (Km^2)	Agriculture Area (Km^2)	Land Area (Km^2)	BPA without BL (Km^2)
1972	Area	60.35	71.32	188.29	319.97
	%	18.86%	22.29%	58.85%	
1984	Area	39.20	79.99	240.58	359.77
	%	10.90%	22.23%	66.87%	
1999	Area	63.87	198.45	156.67	418.99
	%	15.24%	47.36%	37.39%	
2001	Area	128.53	135.71	161.35	425.59
	%	30.20%	31.89%	37.91%	
2003	Area	110.18	125.14	193.60	428.91
	%	25.69%	29.18%	45.14%	
2005	Area	114.31	111.48	205.56	431.35
	%	26.50%	25.85%	47.65%	
2013	Area	138.62	120.55	178.43	437.60
	%	31.68%	27.55%	40.77%	

Table (3): Areas of the Resulted Classes of BPA without BL Using Unsupervised Classification.

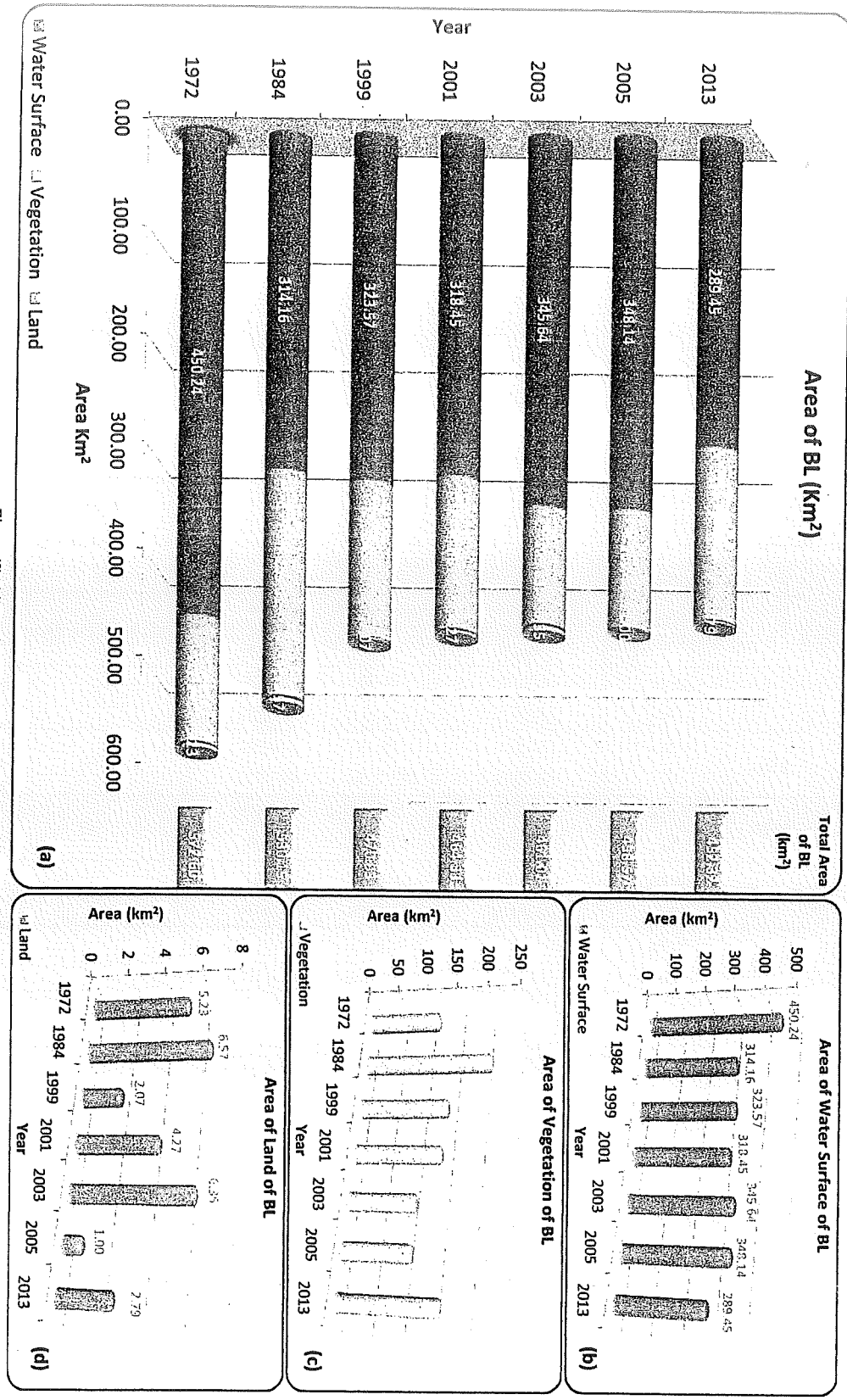


Figure (8): (a) Total Areas of BL from Year 1972 to 2013.
(b) Area of Water Surface Class of BL (c) Area of Vegetation Class of BL (d) Area of Land Class of BL.

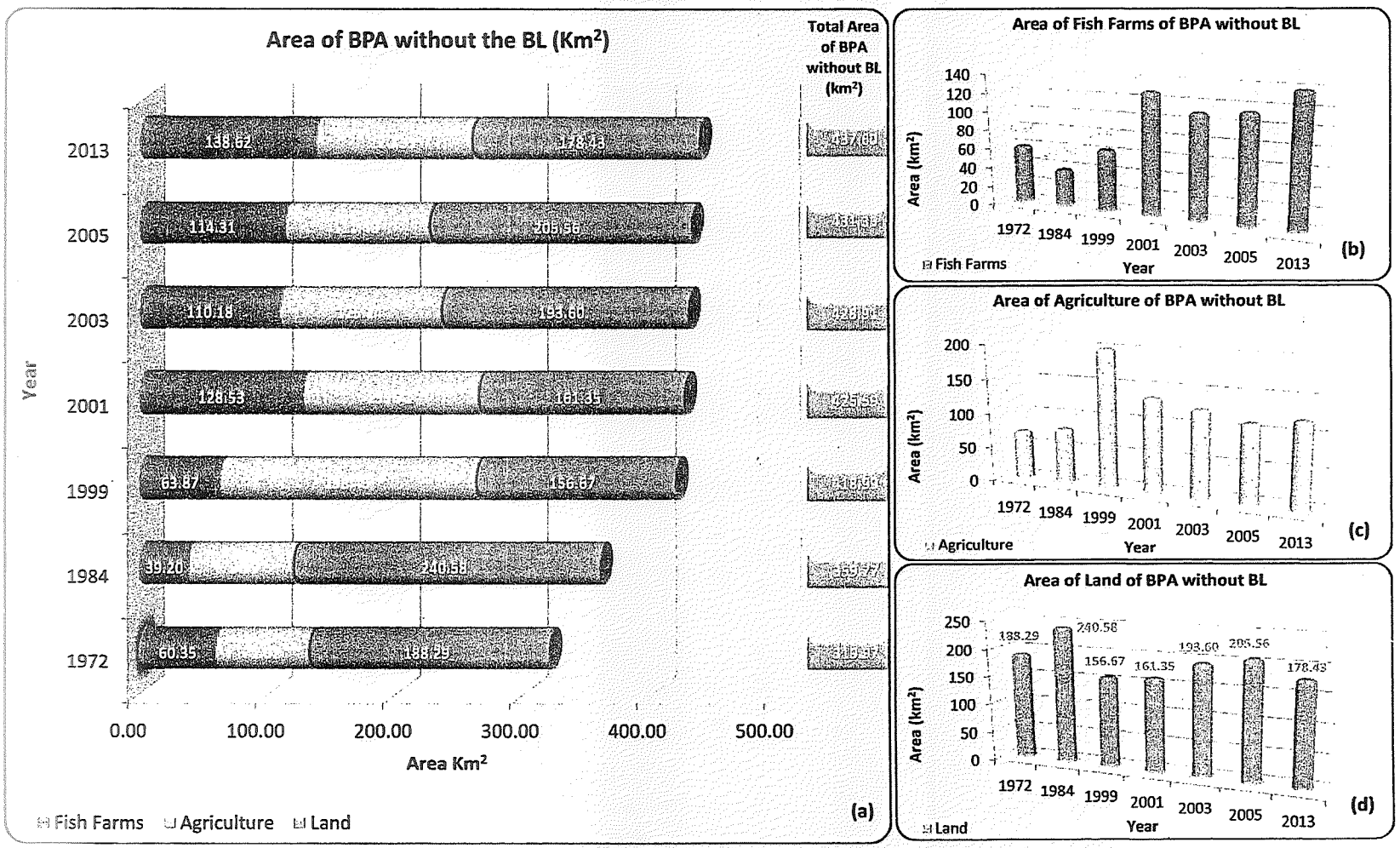


Figure (9): (a) Total Areas of BPA from Year 1972 to 2013.
(b) Area of Fish Farms Class of BPA (c) Area of Agriculture Class of BPA (d) Area of Land Class of BPA.

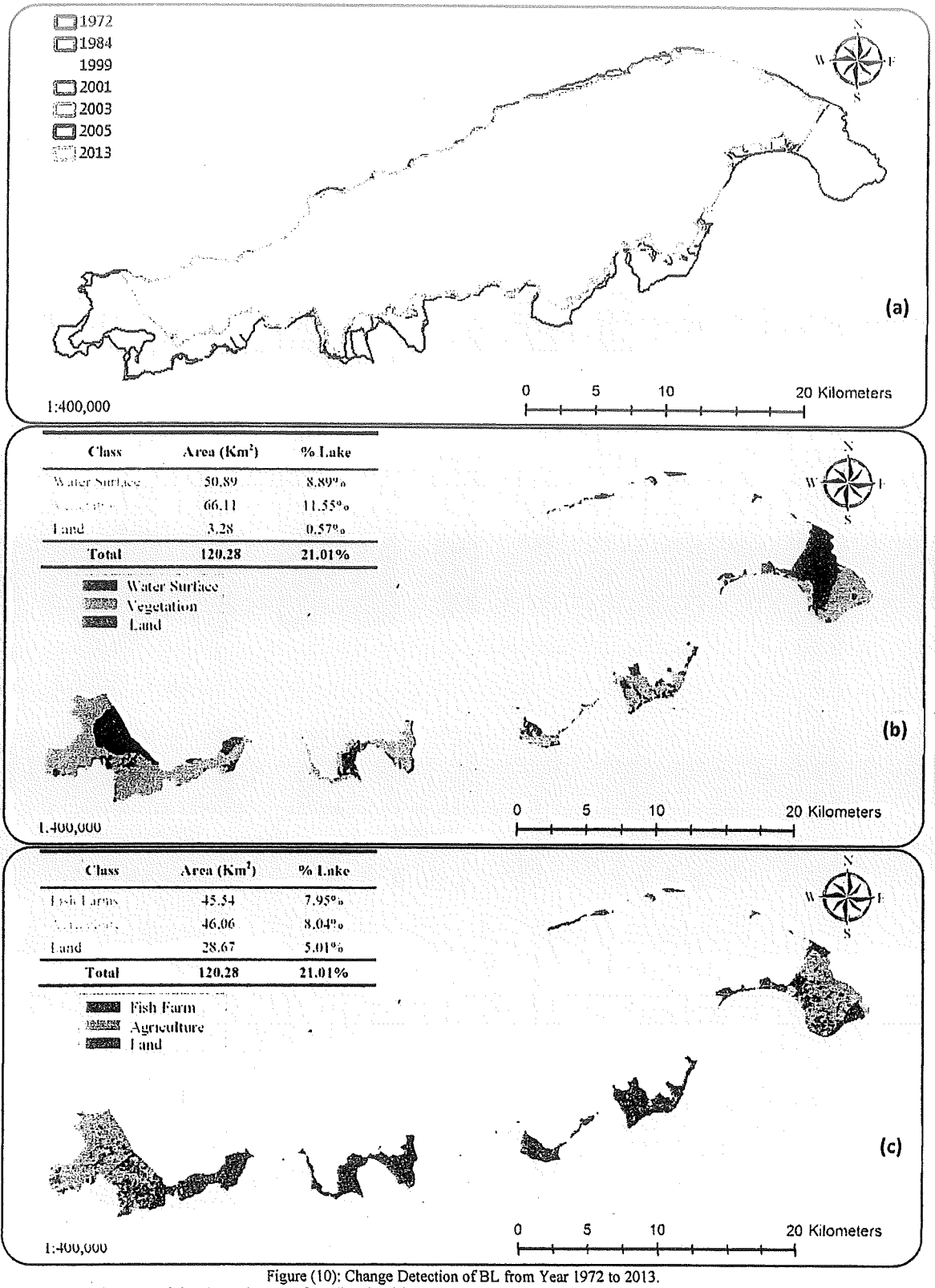


Figure (10): Change Detection of BL from Year 1972 to 2013.
(a) Overlay Map of the Vector Images of BL (b) Classified Image of BL Year 1972 Showing the Changed Part of BL (c) Classified Image of BL Year 2013 Showing the Changed Part of BL.

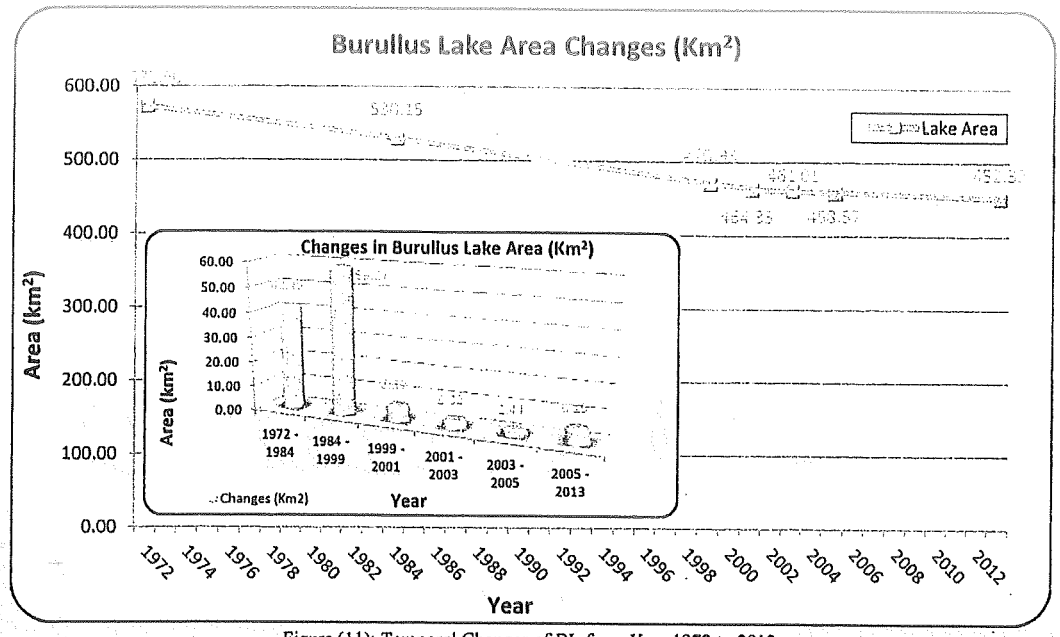


Figure (11): Temporal Changes of BL from Year 1972 to 2013.

Year	Changes in BL Area (Km²)
1972 - 1984	42.45
1984 - 1999	59.21
1999 - 2001	6.60
2001 - 2003	3.32
2003 - 2005	2.44
2005 - 2013	6.25
Total	120.28

Table (4): Temporal Changes of BL According to On-Screen Digitizing of Landsat Imageries.

On the other hand, changes in classes of TMLSI of BPA without the BL elaborate the effect of human activities over the study period, where the blue color refers to fish farming activities, green color to agricultural and land reclamation activities and red color to urban activities. They also indicate that, there is an obvious increase in fish farming from 39.2 km² (10.90%) on 1984 to 138.62 km² (31.68%) on 2013 (total amount of increased fish farming 99.42 km² (10.88%) compared to area of the BPA). In addition, there is an irregular pattern in agriculture from 79.99 km² on 1984 to 198.45 km² on 1999 (mean area coverage over the study period 120.38 km²) and an irregular pattern in urban from 156.67 km² on 1999 to 240.58 km² on 1984 (mean area coverage over the study period 189.21 km²).

In fact, BPA is about 913.92 km². BL was representing 58.01% of BPA at the beginning of its declaration on 1984 as a conservation site while it becomes now 49.49% of BPA on 2013. Finally, there is a noticeable decreasing in water surface of BL at the expense of the vegetation. On the other hand, a great increasing in fish farming outside BL within BPA is at the expense of the agriculture and bare land. These changes

caused by irresponsible fish farming and as well as bridging BL for cultivation since the 1970s.

5. Conclusion

Change detection has become a useful tool for detecting land cover changes from remotely sensed imagery. It enables resource managers to observe changes over large areas and provided long term monitoring capabilities. Generally, we can conclude that digital change detection techniques using temporal multispectral remote sensing data are useful to assist human analysts of remote sensing data, and provide detailed information for detecting and assessing land cover and land use dynamics in the context of spatial, quantitative and statistic themes.

The obtained results from change detection of the TMLSI of the study area of BPA showed that there are noticeable and clear changes in the land use and land cover of BL for the period 1972-2013. BL decreases by 21.01% from 1972 to 2013. The part that was taken from BL is converted to fish farms, agriculture and urban. BPA is exposed to extinction as a result of backfilling BL for land reclamation and fish farming. In addition to the irresponsible fish farming that lies within BL itself and was observed from field visits, and not explicitly mentioned in this paper and needs extra work and analysis based on remote sensing.

The study indicated also that there was a trend change in the location and area of vegetation distribution within BL. In 1972, the majority of vegetation assemblages were located at the southern west part of BL while during the last 29 years there was a shift in location and area of the vegetation toward the central and southern east part of BL. This change is mainly attributed to the reclamation and the drying processes of the fishponds in both of the south east of BPA and at the sand bar of Mediterranean Sea in addition to the increase in the land reclamation areas.

The conclusion of the present study is that the natural resources within BPA are under several pressures and threats from human developmental activities, which may lead to environmental degradation toward their ecosystem tipping points. In addition, the coastal sand dunes and coastal ecosystem require immediate management intervention from the protected area management unit to mitigate the increased human activities (especially after 25th of January Egypt Revolution) seeking for sustainable use of its natural recourses.

6. Recommendations

It is recommended that experts of environment, conservation, resource management and sustainability, ecology, biodiversity, eco-system and inland water explain why these changes happened, in order to save one of the most important wetland in Egypt. They are also required to develop a plan for sustainable use of the site to ensure survival of one of the most relevant eco-system for biodiversity for the next generations.

Acknowledgment

Deep gratitude to Prof. Ibrahim Farag (Professor of Computer Science, Department of Computer Science, Faculty of Computers and Information System, Cairo University) for his support and encouragement. Cordial thanks also are due to Dr.

Khaled Allam (National Project Manager, National Biodiversity Strategy and Action Plan Project "NBSAP Project", GEF/UNDP), Dr. Mohamed Sameh (Head of Geology Department, Nature Conservation Sector, Egyptian Environmental Affairs Agency, EEAA/NCS) and Dr Yousria Hamed (Project Manager, Support to Egypt Protected Areas Project, "SEPA Project", EIECP/UNDP) for their support, kind help, precious remarks, and guidance.

References

[1] Saleh SAH, "Temporal Change Detection of AL-Hammar Marsh–IRAQ Using Remote Sensing Techniques", 1:333, DOI:10.4172/scientificreports.333, 2012.

[2] D. LU, P. MAUSEL, E. BRONDIZIO and E. MORAN, "Change Detection Techniques", International Journal of Remote Sensing (INT. J. REMOTE SENSING), VOL. 25, NO. 12, DOI: 10.1080/0143116031000139863, ISSN 0143-1161 print/ISSN 1366-5901, 2003, p. 2365–2407.

[3] Kassas, M. *et al.*, "Management Plan for Burullus Protected Area", Med Wet Coast, Global Environment Facility & Egyptian Environment Affairs Agency, Cairo, Egypt, 2002, p. 6-12.

[4] G. Zalidis, A. Mantzavelas and E.N. Fitoka, "Mediterranean Wetland Inventory Photointerpretation and Cartographic Conventions", MedWet, Greek Biotope, Wetland Center (EKBY), Instituto da Conservaco da Natureza, Wetland International Publication, VOL. IV, Portugal, ISBN 972-8083-75-0, 1996.

[5] Kamal H. Shaltout, Magdy T. Khalil, "Lake Burullus (Burullus Protected Area)", Egyptian Environmental Affairs Agency, Department of Nature Protection, MedWetCoast Project, Publication of National Biodiversity Unit, No. 13, 2005.

[6] "The First Egyptian-International Conference on Protected Areas and Sustainable Development", Ministry of State for Environmental Affairs, Egyptian Environmental Affairs Agency, Sharm El-Sheikh, Egypt, October 23-26, 2002.

[7] Campbell, "*Introduction to Remote Sensing*", CORINE Land Cover Technical Guide, European Commission, Luxemburg, 2002, pp. 21-53.

[8] A. Akgün, A. Eronat and N. Turk, "Comparing Different Satellite Image Classification Methods: An Application In Ayvalýk District, Western Turkey," XX th International Congress for Photogrammetry and Remote Sensing, Proceedings, 2004.

[9] ERDAS Field Guide, Norcross: ERDAS, Inc., December 2010, pp. 566, 842-844.

[10] Kamal H. Shaltout and Yassin M. Al-Sodany, "Vegetation analysis of Burullus Wetland: a RAMSAR site in Egypt", ©Springer Science+Business Media B.V., Wetlands Ecol Manage, DOI 10.1007/s11273-008-9079-5, 3 January 2008.

[11] P. COPPIN*, I. JONCKHEERE, K. NACKAERTS, B. MUYS, "Digital Change Detection Methods in Ecosystem Monitoring: a Review", Geomatics and Forest Engineering Group, Department of Land Management, Katholieke Universiteit Leuven, Vital Decosterstraat 102, B-3000 Leuven, Belgium and E. LAMBIN, 2004.

[12] Shyam Boriah, Vipin Kumar, Michael Steinbach, Christopher Potter, and Steven Klooster, "Land Cover Change Detection using Data Mining Techniques", TR 08-009, 2008.

[13] Yan Qin, Zhifeng Yang, Wei Yang, "Ecological risk assessment for water scarcity in China's Yellow River Delta Wetland", VOL 25, Issue 5, July 2011, pp. 697-711.

[14] D. Lu, E. Moran, S. Hetrick, and G. Li., "Land-Use and Land-Cover Change Detection", ACT Publication No. 11-02, Ch. 11, 2004.

[15] MEA, Millennium Ecosystem Assessment. "Ecosystems and Human Well-being: Synthesis.Washington", (DC): Island Press, 2005.

[16] Revenga C, Campbell I, Abell R, de Villiers P, Bryer M., "Prospects for Monitoring Freshwater Ecosystems Towards the 2010 Targets", Philosophical Transactions of the Royal Society, B 360, p. 397–413, 2005.

[17] Elsokkary, I.H., "Synopsis on Contamination of the Agricultural Ecosystem by Trace Elements", Egyptian Journal of Soil Science, VOL 36, p. 1-22, 1996.

[18] Hesham M. El-Asmar, Mohamed E. Hereher and Sameh B. El Kafrawy, "Surface area Change Detection of the Burullus Lagoon, North of the Nile Delta, Egypt, Using Water Indices: A Remote Sensing Approach", Received: 02 January 2013, Revised: 17 April 2013, Accepted: 19 April 2013.

[19] H. Farouk, "Coastal Zone Change Detection Using Satellite Images", Master Thesis, Univesity of Helwan, Egypt, Ch 4, p. 118-119, 2005.

[20] H.Farouk and B.M.Abdel Latif, "Classification Based Approach for Spectral Signature of Remotely Sensed Temporal Data", IJCSNS International Journal of Computer Science and Network Security, VOL.13, No.12, December 2013.

[21] Sheikh Saeed Ahmad and Summra Erum, "Remote Sensing and GIS Application in Wetland Change Analysis: Case Study of Kallar Kahar", Department of Environmental Sciences, Fatima Jinnah Women University, Rawalpindi, Pakistan. Sci., Tech. and Dev., 31 (3): 251-259, 2012.

[22] Mohamed E. Hereher; Mahmoud I. Salem and Dina H. Darwish. Mapping water quality of Burullus Lagoon using remote sensing and geographic information system. Journal of American Science 2010; 7(1): 138-143. (ISSN: 1545-1003).

Complexity Reduction of fuzzy Rule Based System: Data-Sensitive Fuzzy Set Approach

Baha Eldin A. Amir, Ahmad M. Gadallah, Hesham A. Hefny

Abstract

A fuzzy system in form of Fuzzy Rule Based Systems (FRBS) is a classic example for both quantitative and qualitative aspects of complexity such as uncertainty of environment, linguistic modeling, approximate reasoning capabilities and processing time. Using well-defined fuzzy sets that is more accurate with specific data usually leads to increase complexity of the fuzzy system. This paper introduces a new simplifying FRBS approach using data-sensitive fuzzy set technique. The proposed approach is a technique that depends on Data-Sensitive fuzzy sets instead, which enables to reduce the complexity of the fuzzy system in both quantitative and qualitative aspects.

This paper is organized in 6 sections. First section introduction, second section the relation between number of rules and the complexity, third section reviews most of existing rule base reduction methods, fourth section discusses the proposed data-sensitive approach, fifth section a case study and finally sixth section results and conclusion.

Keywords

Fuzzy Rule Based System, Complexity, Linguistic Values and data-sensitive.

1. Introduction

Although fuzzy logic is an advanced extension to binary logic since fuzzy logic is cope with situations of vague and imprecise, but binary logic is still used predominately. The main reason for that is the high complexity associated with the use of fuzzy logic. So many of recent researchers are focus on studying and solving complexity problem by means of enhanced approaches to control and minimize complexity which allow us using fuzzy logic more effectively and more efficiently [1] .

The common type of FRBS is Mamdani system which represented by the if-then rules as following:

If i_1 is $v_{i1,1}$ and/or ... and/or i_m is $v_{im,1}$ then o_1 is $v_{o1,1}$ also ... also o_n is $v_{on,1}$

and/or

.....

and/or

If i_1 is $v_{i1,r}$ and/or ... and/or i_m is $v_{im,r}$ then o_1 is $v_{o1,r}$ also ... also o_n is $v_{on,r}$

Also Sugeno System is another FRBS which represented as following:

If i_1 is $v_{i1,1}$ and/or ... and/or i_m is $v_{im,1}$ then $o_1 = f_{1,1}(i_1...i_m)$ also ... also $o_n = f_{n,1}(i_1...i_m)$

and/or

.....

and/or

If i_1 is $v_{i1,r}$ and/or ... and/or i_m is $v_{im,r}$ then $o_1 = f_{1,r}(i_1...i_m)$ also ... also $o_n = f_{n,r}(i_1...i_m)$

Where m is the number of inputs, n is the number of outputs and r is the number of fuzzy rules in the system [1].

Department of Computer and Information Science, Institute of Statistical Studies and Researches, Cairo University

FRBS`s come with two contradictory requirements [2, 3]:

- *Accuracy:* is the capability to faithfully represent the real system through well-defined measures that are widely accepted.
- *Interpretability:* is the capacity to express the behavior of the real system in understandable way. This related to several factors, mainly the model structure, the number of input variables, the number of fuzzy rules, the number of linguistic terms, the shape of fuzzy sets, etc. there is still no standard measure to assess how good interpretability is.

Therefore, it is important when modeling the fuzzy system to pay attention to both aspects.

The associated fuzzy operations in the fuzzification, the inference and the defuzzification stages [3] increase the level of quantitative complexity. This increase of quantitative complexity becomes much more as the number of inputs in fuzzy system gets bigger because the amount of fuzzy operations is a function of the number of fuzzy rules which on its turn depends on the number of inputs. A part of increased quantitative complexity the transparency and the interpretability of fuzzy systems reduced as the number of fuzzy rules increases. In this case it is hard to explain what happening in the system which mean that the problem has been moved from the environment that the fuzzy system is suppose to model , to the system itself [1].

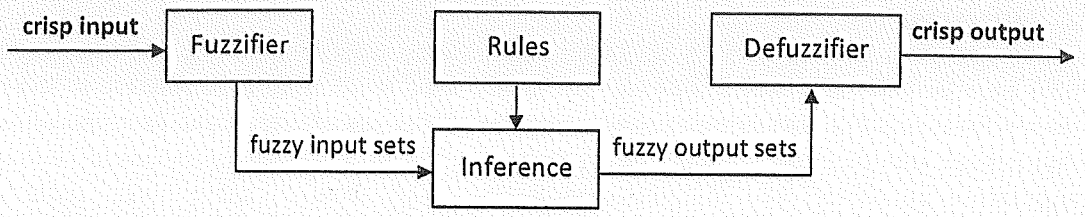


Fig. 1 FRBS

2. *Number of rules and complexity reduction*

Researches in FRBS mainly focused on quantitative complexity issues through methodologies that improve the efficiency and increasing productivity of the fuzzy system. These methodologies are concerning primarily about reducing the time for completion of the required computations by reducing the number of rules.

The maximum number of rules in a fuzzy system r is an exponential function of the number of inputs m and the number of linguistic values w in the form:

$$r = w^m \quad \text{..... Eq. 1}$$

If the number of linguistic values per input is not a constant, then the maximum number of rules in a fuzzy system is given by the following arithmetic product :

$$r = w_1 \dots w_m \quad \text{..... Eq. 2}$$

Fig. 2 showing the impact of number of inputs or number of linguistic values per input would be on the number of rules. It is obvious that the resulting quantitative complexity has to be taken seriously taking in mind that many real life systems are usually much larger number of inputs and often have to be operated in real-time. Beside this, as the number of rules gets bigger it usually harder to understand what is happening in the fuzzy system in other words the

level of qualitative complexity in the fuzzy system increases with the increase of the number of rules. It is important to know that the exact level of this complexity is a function of the overall amount of fuzzy operations during fuzzification, inference and difuzzification, and this amount itself depends on the number of rules.

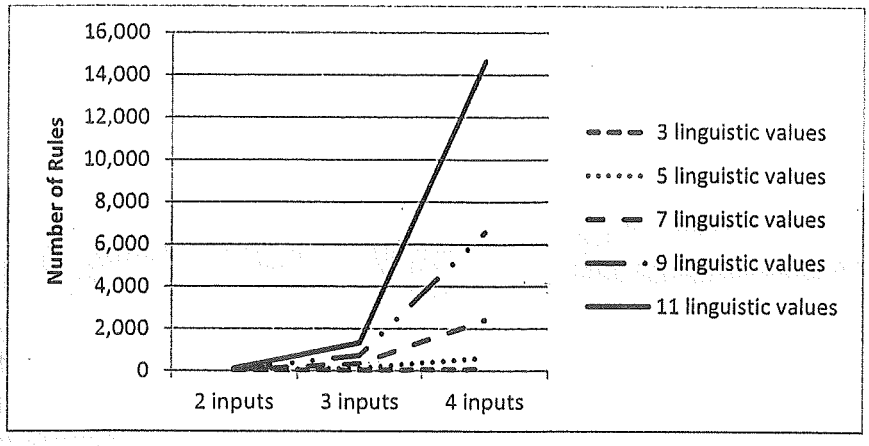


Fig. 2 Number of rules as exponential function of number of inputs

3. Classical methodologies for complexity reduction.

Most of the known methods for complexity reduction in fuzzy systems reduce the number of fuzzy rules by either reducing the number of inputs or the number of linguistic values that these inputs can take. These methods are classified into several groups as following [6, 7, 8, 9, 10, 11, 12]:

3.1 Removing less significant and/or merging similar linguistic values.

This first group is based on removing less significant and/or merging similar linguistic values.

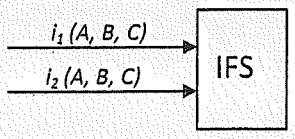


Fig. 3 Initial Fuzzy System

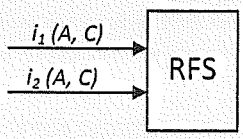


Fig. 4 Reduced Fuzzy System

The initial Fuzzy System (IFS) is described by the two inputs i_1 and i_2, and each of these inputs can take the three linguistic values *A, B, C.* In the RFS, the linguistic value *B* removed in which *A, C* will cover all crisp input values).

Table 1 Initial Fuzzy System

Rule number	Linguistic value of i_1	Linguistic value of i_4
1	1	1
2	1	2
3	1	3
4	2	1
5	2	2
6	2	3
7	3	1
8	3	2
9	3	3

Table 2 Reduced Fuzzy System

Rule number	Linguistic value of i_1	Linguistic value of i_4
1	1	1
2	1	2
3	2	1
4	2	2

3.2 Removal less significant /or fusing similar inputs.

The second group of methods for rule base reduction based on removing less significant and/or merging similar linguistic values.

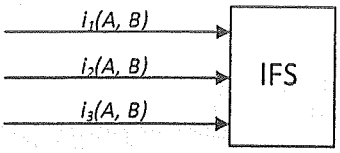


Fig. 5 Initial Fuzzy System

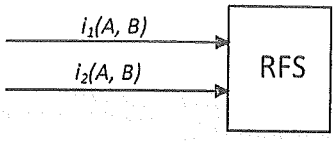


Fig. 6 Reduced Fuzzy System after removal of input i_3

Table 3 Initial Fuzzy System

Rule number	Linguistic value of i_1	Linguistic value of i_2	Linguistic value of i_3
1	1	1	1
2	1	1	2
3	1	2	1
4	1	2	2
5	2	1	1
6	2	1	2
7	2	2	1
8	2	2	2

Table 4 Reduced Fuzzy System

Rule number	Linguistic value of i_1	Linguistic value of i_2
1	1	1
2	1	2
3	2	1
4	2	2

Table 5 Reduced Fuzzy System after fusion of inputs i_2 and i_3 into i_4

Rule number	Linguistic value of i_1	Linguistic value of i_4
1	1	1
2	1	2
3	2	1
4	2	2

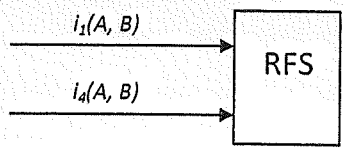


Fig. 7 RFS after fusion of inputs i_2 and i_3 into i_4

3.3 Singular value decomposition of the matrix

The third group of methods is singular value decomposition of the matrix based on representing the crisp values of the output. For example, if the IFS is described by the two inputs i_1, i_2 and each of these inputs can take 5 linguistic values negative big (NB), negative small (NS), zero (Z), positive small (PS) and positive big (PB). So if NB=1, NS=2, Z=3, PS=4 and PB=5 in the IFS, and if N=6 and P=7 in the RFS.

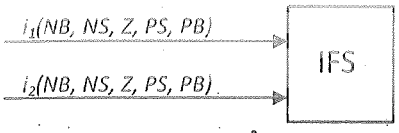


Fig. 8 Initial Fuzzy System with $5^2 = 25$ rules that covers all possible permutations of inputs.

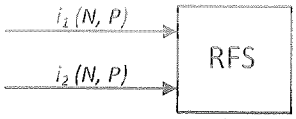


Fig. 9 Reduced Fuzzy System

Table 6 Reduced Fuzzy System by singular value decomposition with linguistic values N=6 and P=7

Rule number	Linguistic value of i_1	Linguistic value of i_2
1	6	6
2	6	7
3	7	6
4	7	7

3.4 Conversion into Union Rule Configuration

The fourth group of methods based on converting the intersection rule configuration of a fuzzy system into a union rule configuration with a smaller number of rules.

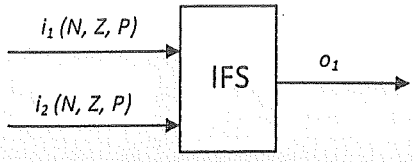


Fig. 10 Initial Fuzzy System

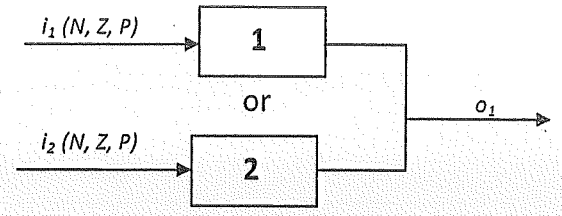


Fig. 11 Reduced Fuzzy System

Table 7 Initial Fuzzy System

Rule number	Linguistic value of i_1	Linguistic value of i_2	Linguistic value of o_1
1	1	1	$v_{o1,1}$
2	1	2	$v_{o1,2}$
3	1	3	$v_{o1,3}$
4	2	1	$v_{o1,4}$
5	2	2	$v_{o1,5}$
6	2	3	$v_{o1,6}$
7	3	1	$v_{o1,7}$
8	3	2	$v_{o1,8}$
9	3	3	$v_{o1,9}$

Table 8 First term in the reduced Fuzzy System with union rule configuration

Rule number	Linguistic value of i_1	Linguistic value of i_2
1	1	$v_{o1,1}$
2	2	$v_{o1,2}$
3	3	$v_{o1,3}$

Table 9 Second term in the reduced Fuzzy System with union rule configuration

Rule number	Linguistic value of i_1	Linguistic value of i_2
1	1	$v_{o1,1}$
2	2	$v_{o1,2}$
3	3	$v_{o1,3}$

3.5 Spatial Decomposition into Subsystems

The fifth group of methods converts a fuzzy system into spatially decomposed subsystems as a result the overall number of rules is reduced.

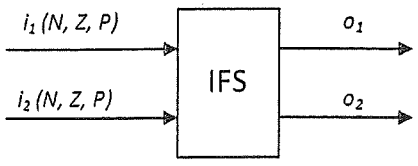


Fig. 12 Initial Fuzzy System

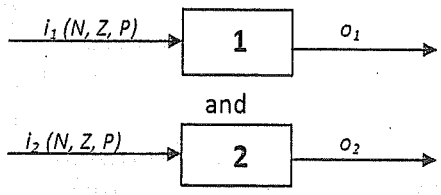


Fig. 13 Reduced Fuzzy System

Table 10 Initial Fuzzy System

Rule number	Linguistic value of i_1	Linguistic value of i_2	Linguistic value of o_1	Linguistic value of o_2
1	1	1	$V_{o1,1}$	$V_{o2,1}$
2	1	2	$V_{o1,2}$	$V_{o2,2}$
3	1	3	$V_{o1,3}$	$V_{o2,3}$
4	2	1	$V_{o1,4}$	$V_{o2,4}$
5	2	2	$V_{o1,5}$	$V_{o2,5}$
6	2	3	$V_{o1,6}$	$V_{o2,6}$
7	3	1	$V_{o1,7}$	$V_{o2,7}$
8	3	2	$V_{o1,8}$	$V_{o2,8}$
9	3	3	$V_{o1,9}$	$V_{o2,9}$

Table 11 First subsystem in the reduced Fuzzy System

Rule number	Linguistic value of i_1	Linguistic value of o_2
1	1	$V_{o1,1}$
2	2	$V_{o1,2}$
3	3	$V_{o1,3}$

Table 12 Second subsystem in the reduced Fuzzy System

Rule number	Linguistic value of i_2	Linguistic value of o_2
1	1	$V_{o1,1}$
2	2	$V_{o1,2}$
3	3	$V_{o1,3}$

3.6 Decomposition into Multilayer Hierarchical Structure

Sixth group of methods rearranges the inputs in a fuzzy system in a way that leads to the reduction of the number of rules.

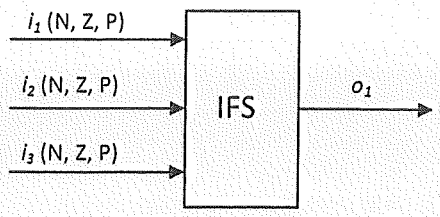


Fig. 13 Initial Fuzzy System with $3^3 = 27$ rules

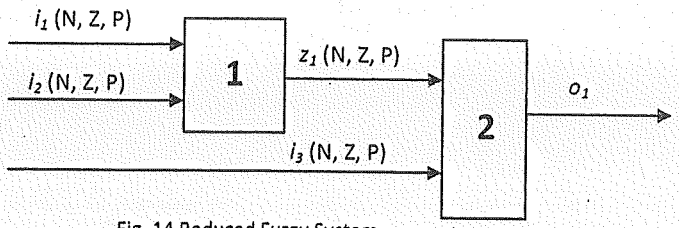


Fig. 14 Reduced Fuzzy System

Table 13 First layer in the reduced Fuzzy System with Multilayer Hierarchical Structure

Rule number	Linguistic value of I_1	Linguistic value of i_2	Linguistic value of z_1
1	1	1	$V_{z1,1}$
2	1	2	$V_{z1,2}$
3	1	3	$V_{z1,3}$
4	2	1	$V_{z1,4}$
5	2	2	$V_{z1,5}$
6	2	3	$V_{z1,6}$
7	3	1	$V_{z1,7}$
8	3	2	$V_{z1,8}$
9	3	3	$V_{z1,9}$

Table 14 Second layer in the reduced Fuzzy System with Multilayer Hierarchical Structure

Rule number	Linguistic value of z_1	Linguistic value of i_1	Linguistic value of o_1
1	1	1	$V_{z1,1}$
2	1	2	$V_{z1,2}$
3	1	3	$V_{z1,3}$
4	2	1	$V_{z1,4}$
5	2	2	$V_{z1,5}$
6	2	3	$V_{z1,6}$
7	3	1	$V_{z1,7}$
8	3	2	$V_{z1,8}$
9	3	3	$V_{z1,9}$

Most of the above methods for rule base reduction in fuzzy systems have some serious drawbacks [1] such as:

- Limited scope.
- Approximate behavior of the IFS.
- Empirical nature.

The process of these methods usually associated with loss or aggregation of information that results that RFS may not represent adequately the IFS.

4. Proposed approach.

Proposed approach considers two important issues about FRBS:

- Fuzzy sets are the core element in Fuzzy Logic, they are characterized by membership functions which is associated with terms or words used in the antecedents and consequents of rules with input and output to the fuzzy system. In real world, many of fuzzy systems modeled with specific and accurate fuzzy sets to improve transparency and interpretability of the system. This interpretability comes with high price of quantitative complexity according to increasing of number of rules dramatically as shown in Fig. 2.
- Complete and consistent fuzzy rule base systems are very desirable not only because it provides detailed information about the fuzzy system by means of all possible permutations of linguistic values of inputs but also because its inference mechanism is non-contradictory. These properties are related to the antecedents part of the rule base [1].

The proposed approach is targeting complete and consistent FRBS through data-sensitive simplification technique. This technique may differ in detail from FRBS to another, so we must practice and examine data of each single system to recognize data properties of the system in order to determine the appropriate technique to be used.

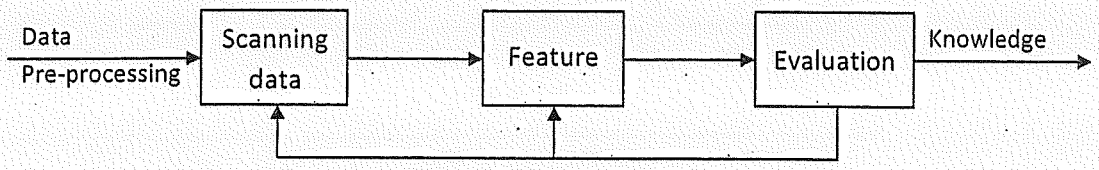


Fig. 15 Data-sensitivity practicing flow.

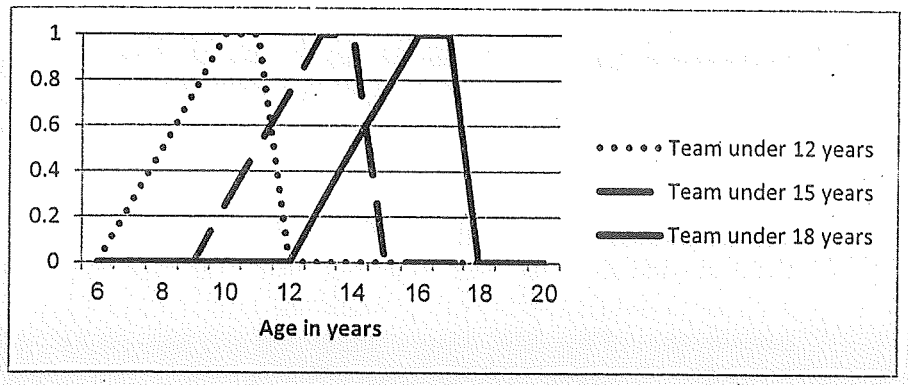
As shown in Fig 15 the knowledge that figured out from the previous process represents the properties of data of the FRBS that based on the relations between objectives of data for example:

1- Linguistic values categories.
2- Linguistic values that frequently used in antecedent part of the rule.
3- Linguistic values average.

This knowledge is used to re-design the fuzzy sets that will represent the simplified system as illustrated in the following section.

5. Case Study.

The case study that illustrating the data-sensitive approach is about a modeled FRBS which is concerned with selecting candidates that their heights are satisfying basketball team among students of schools. This fuzzy system is described by 4 inputs i_1, i_2, i_3 and i_4, first input i_1 represents the age, second input i_2 represents Prim_Height, third input i_3 represents Prep_Height and fourth input i_4 represents Secondary_Height, each input can take 3 linguistic values. The initial fuzzy system presented as below:



.Fig. 16 Age membership function

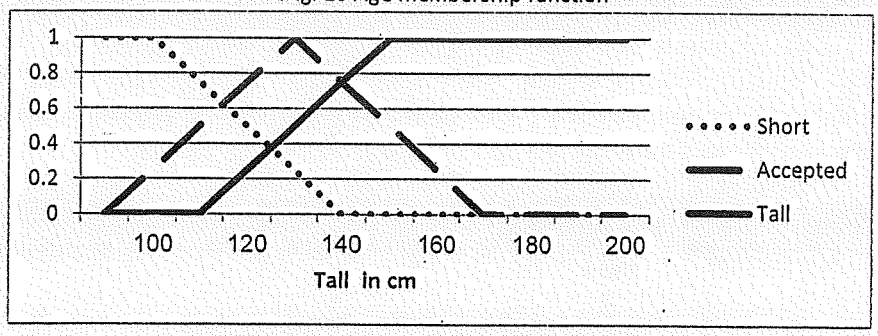


Fig. 17 Primary school height membership function

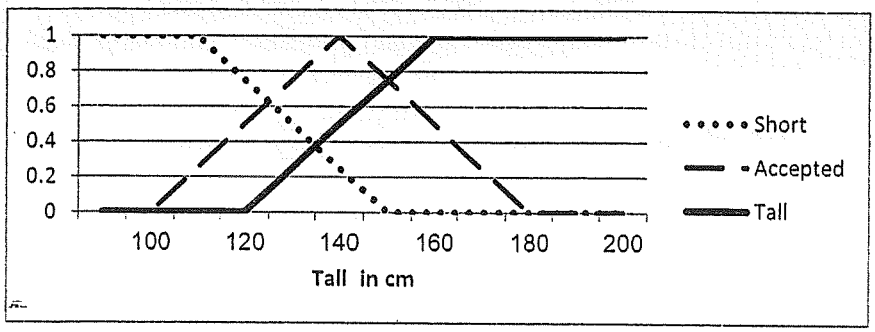


Fig. 18 Prep. school height membership function

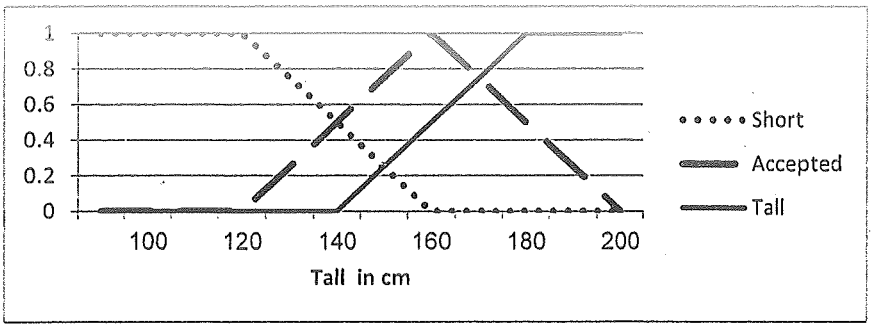


Fig. 19 Secondary school height membership function

Table 15 Initial Fuzzy System

Rule number	Linguistic value of i_1	Linguistic value of i_2
1	1	1
2	1	2
3	1	3
4	1	4
5	1	5
6	1	6
7	1	7
8	1	8
9	1	9

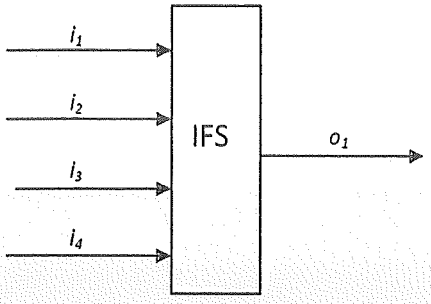


Fig. 20 Initial Fuzzy System

According to the above case there are four fuzzy variables, each has three linguistic values. So according to Eq.1 and Eq.2 the number of rules in the initial fuzzy system will be 3^4 which is 81 rules as complete and consistent system or 3 initial subsystems for each class of school with 3 rules as shown in table 15.

According to data-sensitivity approach we practice and examine the data of the initial fuzzy system as in Fig. 15 and the knowledge figured proposes a simple technique that uses a function of age that assumes a one fuzzy set with three sensitive or dynamic linguistic values as following:

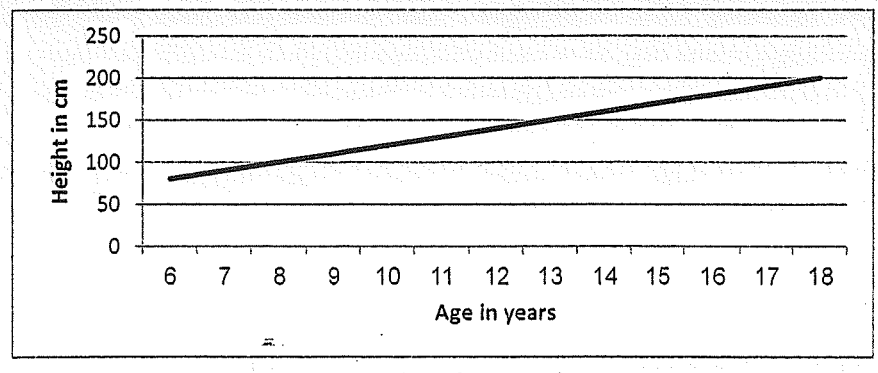


Fig. 21 Height as function of age

From the above chart, we can conclude that the age of a specific person determine the shape of the predicted data-sensitive fuzzy set meanwhile the average height of a specific age becomes the core of the fuzzy value "Medium_Heigt". For example if we select age of 13 then the average height for this age is 1.5 meters, which represent the core of "Accepted" of the assumed fuzzy set for this age.

The reduced fuzzy system will be: If i_1 is fn(age). Which is mean all the time the system uses only one dynamic fuzzy set according to this function instead of three fuzzy sets to cover the three desired teams. The proposed fuzzy se will be as following:

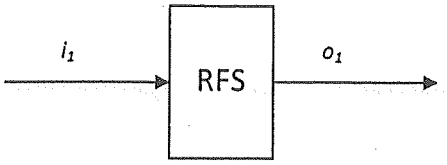


Fig. 23 Reduced Fuzzy System

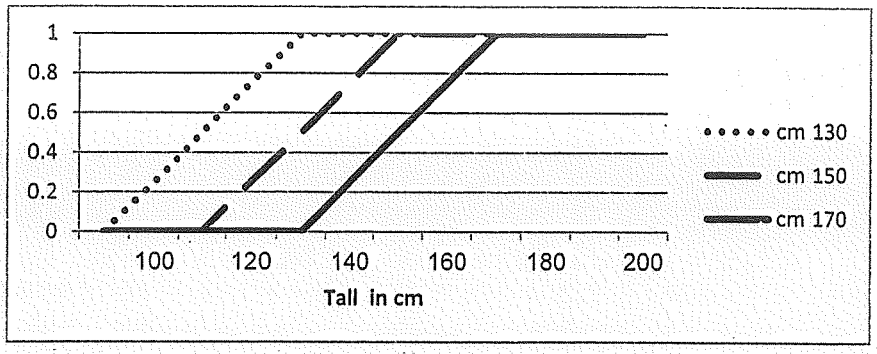


Fig. 22 Data-Sensitive height membership function.

6. Conclusion.

The proposed technique presents a more flixable manner for analysing and synthesising a more simplified FRBS. In particular, data-sensitive technique could generalize many of real world applications that have been designed in a more complicated manner. This new approach has impact mainly on the quantitative complexity. In other words dynamic or general fuzzy set which cover the universe of discores which mean reducing number of linguistic values leads to reduce no of rules which mean reducing processing time through fuzzy system stages (fuzzification, processing and defuzzification). From a wider perspective that takes into account factors that contribute to the qualitative complexity of the fuzzy system.

In addition the new approach may not only be able to reduce the quantitative complexity of the system it will lead systematically to make the system more generalized and widespread through different data types in different environments.

The potential advantages for this approach in comparison with known complexity reduction methods are :

- It would guarantee that the behavior of the reduced fuzzy system is equivalent not approximate to the behavior of the initial fuzzy system.
- It would be applicable to a fuzzy system with any number of inputs/outputs.
- It would use a simple formal model.
- It would be applicable for a variable number and meaninig of the linguistic values of the inputs/outputs.

The limitation with this technique if we hae a non-symetric fuzzy variables. In this case we have to seek another data-type-sensitivity to be able to reduce the complexity of the system.

References

1. Gegov A (2007) Complexity Management in Fuzzy Systems, A rule Base Compression Approach. Studies in Fuzziness and Soft Computing, volume 211.
2. Gacto M.J., Alcalá R., Herrera F. (2011) Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures. Information Sciences 181 (2011) 4340–4360, 2011 Elsevier Inc.
3. Riid A, R¨ustern E (2011) An Integrated Approach for the Identification of Compact, Interpretable and Accurate Fuzzy Rule-Based Classifiers. In: Intelligent Engineering Systems (INES), IEEE international conference p 101-107.
4. Galende M, Sainz G.I, Fuente M.J (2011) Accuracy-Interpretability Trade-Off for Precise Fuzzy Modeling using simple indices. Application to Industrial Plants. International Federation of Automatic Control (IFAC), Preprints of the 18th IFAC World Congress Milano (Italy) August 28 - September 2, 2011 .
5. Galende M, Sainz G.I, Fuente M.J (2011) Complexity reduction and interpretability improvement for fuzzy rule systems based on simple interpretability measures and indices by bi-objective evolutionary rule selection. Springer-Verlag 28 July 2011.
6. Carmona P, Castro J.L. (2008) An Improved ACO Based Plug-in to Enhance the Interpretability of Fuzzy Rule Bases with Exceptions. Advances in Artificial Intelligence Springer-Verlag 2008, Volume 5127 pp 13-24.
7. E. Lughofer E. (2011) : Evolving Fuzzy Systems – Method., Adv. Conc. and Appl., STUDFUZZ 266, Interpretability Issues in EFS pp. 261–291. Springer-Verlag Berlin Heidelberg 2011.
8. Riid A, R¨ustern E (2011) Interpretability, Interpolation and Rule Weights in Linguistic Fuzzy Modeling, Fuzzy Logic and Applications. Springer-Verlag 2011, Volume 6857/2011, DOI: 10.1007/978-3-642-23713-3_12 p 91-98.
9. Mencar C., Castellano G., Fanelli A. M. (2007) Distinguishability quantification of fuzzy sets. Information Science, Volume 177 issue 1 (2007) p: 130-149.
10. Hefny H. A. (2007) Comments on "Distinguishability quantification of fuzzy sets". Information Science, Volume 177 (2007).
11. P. Cordero P., Enciso M., Mora A., de Guzm_an I. P. (2010) A complete logic for fuzzy functional dependencies over t-norms.
12. Sharma D (2011) Designing and Modeling Fuzzy Control Systems, International Journal of Computer Applications (0975 – 8887) Volume 16– No.1, February 2011.

A Survey on Semantic Personalization of E-learning Systems

Mohammed A. Moussa[1] Hesham A. Hefny[2] Akram I. Salah[3]

Abstract:

The issue of personalization in E-learning has become an important topic of research in recent years. With the emergence of Web-based learning systems, it has become possible to provide access to content to wider learner community, and thus offer various learning opportunities for the learners who are not capable to deal with the formal education. Our aim in this paper is to overview various techniques for personalization in the Semantic E-learning domain and choose the technique capable for realizing the Social Collaboration and Communication between the learners and also the adaptability of the learning materials to different learners in various learning environments, as a first step to build a student model through a social network. An ultimate objective in our research is to obtain an enhanced personalized E-learning system that can adapt to students' skills and knowledge, thus each student would learn better and the learning outcome would be maximized.

Keywords:

Adaptive Learning, Collaborative Filtering, Content Filtering, Context, Personalized E-learning systems, Recommendation, Semantic E-Learning, Social Networks, User Profile.

1. Introduction

Traditional teaching methods mainly use a "one size fits all" approach in which the information offer was standardized and equal for everyone. In

1 Msc. student, Institute of Studies and Statistics Researches, Cairo University - Computer Lab. Engineer, Modern Academy for Engineering & Technology in Maadi

2 Head of Department of Computer & Information Sciences, Institute of Studies and Statistics Researches, Cairo University

3 Associate Professor, Department of Computer Sciences, Faculty of Computers & Information, Cairo University

recent times, on the contrary, it has become more clear that different people learn in different ways and that a personalized approach can improve the learning process helping people becoming effective life-long learners **[1].**

Recently, Personalized E-Learning has become an active field of research and experimentation. It represents the Web-based delivery of personalized, comprehensive, dynamic learning contents, aiding the development of communities of knowledge, linking learners and practitioners with experts. In this context, an important role is played by the definition of educational structure that must be contextualized and tailored on the basis of the requirements of two main actors: i) teachers, who have personal teaching approaches, and ii) students, who have personal studying approaches . Moving to this direction, emerging Semantic Web technologies have changed the focus of e-learning systems from task-based approaches to knowledge-based ones, thus the educational systems have been shifted from a teacher- centered perspective to a learner-centered one **[2].**

An adaptive educational system has to provide personalization to the specific needs, knowledge and background of each individual student. Creating an adaptive learning system that meets students' requirements can be challenging since students learn with not only different needs, but also different learning characteristics. A solution to this challenge is the technology of student modeling but its use has been extended to most of the current educational software applications that aim to be adaptive and personalized. Thus a student model is the base for personalization in computer-based educational applications **[3].**

The idea of E-Learning Personalization can be summarized as shown in Fig. 1 **[4].**

Personalization of learning can be expressed as a cycle including a tracking system, a learner's profile, a personalization mechanism, a recommendation database, and an individual learning path. First, Personality tests and personal data deliver a lot of information about the learner representing a static view of the learner's preferred learning way, while the dynamic information about the way of the user's learning is acquired from the tracking system. The tracking system is responsible of monitoring the behavior of the learner during training. The resultant

information is used to create a user profile that is, a certain category of learners. Each category is associated with learning path specific for this profile. This path can refer to various aspects of individual learning as the content of a course, presentation of the content, navigation or collaboration with the system. Finally, the Personalization mechanism - representing the ontology for the content of courses and data mining methods for tracking - is applied in an off-line mode. Through this mode, the system collects learner's data, analyzes them and recommends a teacher changes in the course content.

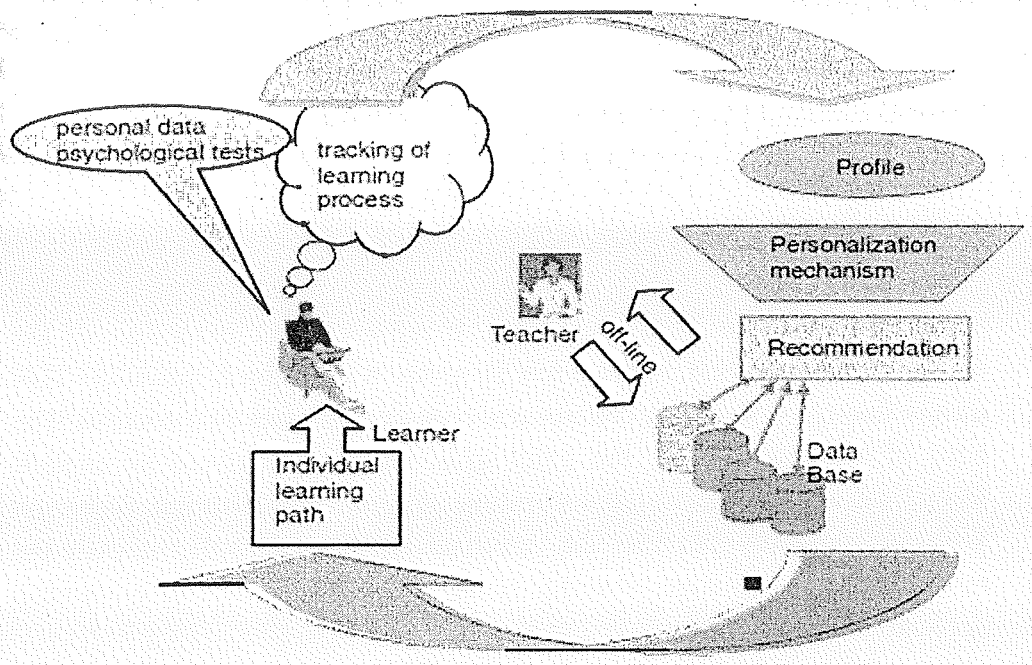


Fig. 1: "The idea of learning personalization in e-learning systems" [4]

This paper is organized as follows:

Section 2 describes the background terminologies used in personalized e-learning systems. Section 3 explains the classification of the approaches for realizing a personalized semantic e-learning system while Section 4 shows a comparison between these approaches. Finally, conclusion is presented in the last section followed by the acknowledgments and references used in this survey.

2. Background

In this section we will discuss the basic terminologies relevant to our field of study - semantic web-based e-learning personalization - as follows:

2.1 E-learning

E-learning (Electronic Learning) stands for all forms of Web-based learning and uses computers and network technologies to create, store, deliver, manage and support online learning courses to anyone, anytime and anywhere. It provides a configurable infrastructure that can integrate learning materials, tools, and services into a single solution for offering training or educational materials quickly, effectively, and economically **[5].**

2.2 The Semantic Web (SW)

Semantic Web is a W3C[4] initiative. It represents an extension of the current Web in which information is given well-defined meaning. This means that the documents contain not only content, but also context. The core of semantic web is the ontologies which are defined as explicit formalization of a shared understanding of a conceptualization. Semantic Web technologies provide the mechanisms for semantic knowledge representation, exchange, sharing, reuse and collaboration of e-learning applications **[6].**

2.3 Semantic E-learning

Semantic E-learning is the e-learning based on the Semantic Web technologies that can easily provide learning materials in a common format and therefore enhance personalized learning. In this context, e-learning systems have the ability of representing semantics of knowledge resources (learning objects, lessons, courses, etc.) and their relationship in a standard format that can permit the dynamic composition of learning materials and processes and enforce the customization by organizing these

---

4 World Wide Web Consortium, http://www.w3.org

resources based on the user's need, such as teacher's preferences or student's profile [7].

2.4 Web Personalization

Web personalization can be defined as any action that customizes the information or services provided by a web site to an individual. When personalization is applied to the semantic web it offers many advantages when compared to the traditional web because semantic web integrates semantics with the unstructured data on web so that intelligent techniques can be applied to get more efficient results. The machine understandable feature of semantic web provides strategies that can perform effective personalization such that the results returned to the user are more relevant to the goal set by him [8].

2.5 Social Networking

Social Networking is a medium that encourages some cooperation and also collaboration which are used in real learning environments like classrooms, boardrooms and online discussion rooms. It also represents a grouping of individuals into specific group with specific interest. In personal E-learning environments, a social network is needed for making the connection between the learners easier, personality changing, solving problems, receiving feedback and getting more expertise [9].

2.6 Recommendation Systems

With respect to the field of E-Learning, Recommendation systems act as an important tool to recommend the best suited and useful piece of learning material for the learner. Usually, the learner faces the problem of spending too much time in searching the desired material but most often ends up in irrelevant (not so important) learning material. High redundancy of the learning material, lack of detail and many other problems are faced by the learner, thus using Recommendation Systems to solve these problems is vital [10].

3. Classification of the approaches for realizing a personalized semantic e-learning system

Through this survey, the Recommendation Systems (discussed in section 2.6) used in personalizing the e-learning systems can be classified into six approaches, as shown in Fig. 2

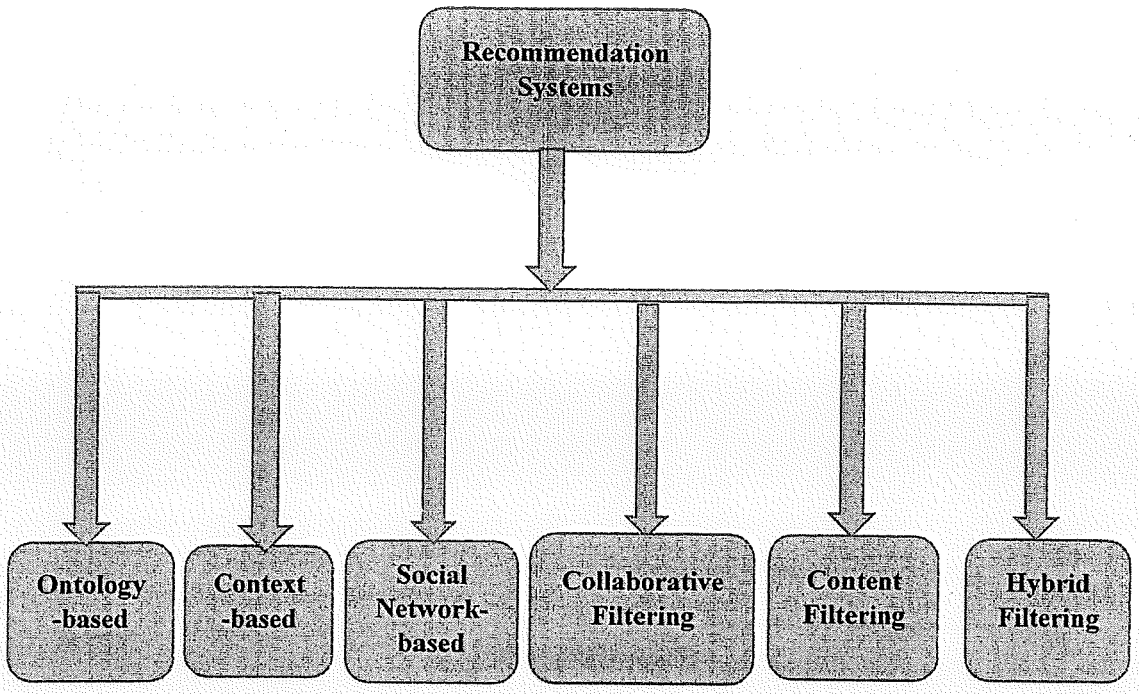


Fig. 2: "Classification of the approaches for realizing a personalized semantic e-learning system"

The approaches and the related work in each are discussed in details in the following sections.

3.1 Ontology-based Approach

An Ontology-based system can be used to personalize the semantic web. This is achieved by using the concepts of domain ontology which contains the information regarding the domain of interest in an ontology format **[11].**

Gaeta, M. et al. [12] proposed an approach named the Intelligent Web Teacher (IWT) describing how personalization can be applied in e-

Learning systems. The IWT model represented an e- Learning platform enabling the definition and the execution of personalized learning experiences, packaged in a Unit of Learning (UoL).
The proposed approach allowed executing more efficient and effective e-Learning processes. Also, the knowledge domain of a course could be easily represented by the e-Learning ontologies. However, the algorithms used were not able to extract concepts from knowledge bases of multiple heterogeneous documents. This could have led to an improvement in the ontology creation for the preparation of personalized e- Learning courses.

3.2 Context-based Approach

Context represents any information that can be used to characterize the situation of an entity where an entity can be a person, place, or object that is considered relevant to the interaction between a user and an application. The user's context is automatically collected by context aware computing which can extract information relevant to the current context and compute a wider range of contextual attributes such as the user's activities, current positions, and their surrounding environments **[13].**

Eyharabide, V. et al. [14] proposed a model named AdaptWeb system to improve e-learning environment personalization. The system focused on making use of a better understanding and modeling of the user's educational and technological context using ontologies.

The proposed model had the advantage of providing the content and navigation recommendations depending on the student's context. However, this model had been applied to academic scenarios but it hadn't been tested in the actual cases

While in **[15], Kurilovas, E. et al.** proposed a swarm-based approach where the learning path depends on the previous learners' decisions. The learning path for personalized learning can be selected using different learners' preferences: their prerequisite knowledge level, learning style, etc. In the presented model, learning styles were chosen as the main criteria.

This model could solve the issue of how the learners' characteristics could be supported in e-learning systems such as Virtual Learning Environments, Adaptive Hypermedia Systems, and Personal Learning Environments. Moreover, the course content (learning objects) is not fixed

(it could be changed over time according to learners' preferences). However, using a dynamic algorithm made this approach more complex than the ontology-based one.

3.3 Social Network-based Approach

Social Network Analysis (SNA) is one of the greatest approaches to classify the user behavior by measuring the intensity of communication among them and also classifying the behavior of student. Social Network Analysis can also be used to investigate the social interactions in classrooms. **[16].**

Xu, Y. et. al [17] proposed an approach named **CSNSA** (Combined Social Network and Semantic Analysis) based on the design and development of a two-layer network analysis framework combining the social network and semantic concept analyses. While Social Network Analysis explores social relationships among researchers, the semantic analysis enabled the extraction of the semantic relationship of the content of researchers' expertise.

The advantages of the proposed network model are represented in transforming the problem of researcher recommendation into a network search problem generating satisfactory recommendation results in addition to supporting researchers' sharing knowledge and communication. However, the proposed approach was dependent only on the degree of interest.

While in **[18], Mansur, A. B. F. et. al** discussed the problems of the previous approach and improved the personalization of the e-learning system through combining the SNA (Social Network Analysis) that only depend on the degree of interest with three meaningful learning clusters: active, constructive and intentional.

This approach led to a better analysis than the previous one that only focused on treating the relationship of the user inside a social network as a node and an edge as it could classify the behavior of students from their log activities in Moodle E-learning system. However, there was a need for implementing a vital cluster describing the collaborative action to give deep analysis to the student's behavior and also a complete description of every student during the e-learning process.

3.4 Collaborative Filtering Approach

The ratings given to a particular resource by the user are aggregated and compared with the ratings given by other users. The similarity measure is calculated between the users. Users with high similarity measure can predict the ratings for one another. The semantic knowledge is incorporated to improve the performance of traditional collaborative filtering recommendations system. While Traditional collaborative filtering system computes exact match to find similar users and items, Semantic collaborative filtering system uses semantic match **[19].**

Kim, H. H. [20] presented a method to provide personalized recommendation services using a tagging ontology for a social e-learning system called TagSES which models relationships among students, lecture materials and tags, then mapping the tags to the domain ontology. Based on the reasoning rules, students with similar interests were clustered and relevant lecture materials were recommended.

Although this approach could realize the learning process in real life for both students and teachers and also used open-source tools like MySQL database system and Elgg: an open source social networking engine. It used PHP for implementing the ontology, though better alternatives could have been used in the ontology management.

While in **[21], Klašnja-Milićević, A.** presented in his Phd. Thesis a model to select tags that can express the preferences and characteristics of users required to generate personalized recommendations and options on the use of models for personalized tutoring system.

This model could help the user organize his or her personal collections as the tagging information could be used to make recommendations. However, it did not focus more specifically on measuring the impact of prior learner experience and interest on the effect of creating tags. Moreover, possible learners could be mood or stress level.

3.5 Content Filtering Approach

Semantic content-based filtering is based on using semantic relations. Semantic content-based filtering can enhance content based recommendation by addressing the Cold start problem appearing during the traditional content-based filtering. Cold start problem occurs when

there are not enough user ratings and the number of available users or resource ratings is insufficient [22].

Cheung, R. et al. [23] proposed a personalized e-learning system (PES) using an adaptive framework, for students to learn basic Java programming, with adaptive features according to the students' learning preference and performance.

This approach provided an easy way to improve the course content. Also, it could easily manage the distributed learning resource with a single learning model for the user. However, this system was not suitable for high school students who do not have online learning experience. It was acceptable for only learners having online learning experience, but want to improve their learning process and effectiveness through an adaptable learning platform.

3.6 Hybrid Filtering Approach

Apart from the above mentioned techniques there are several other techniques used for recommendation systems. These techniques can use a combination of the semantic collaborative-based filtering, the semantic content-based filtering techniques and rules -defined as patterns and sequences- to generate hybrid recommendations that take the advantages of all filtering techniques to produce high quality recommendations for better personalization [24].

Köck, M. et al. [25] proposed an approach named DMM (Discrete Markov Models) to address the extraction of information from sequential learners' activities, and modelling these activities into clusters which can be applied at three levels: predefined problem-solving styles (level I), problem-solving styles along predefined learning dimensions (level II), and learning dimensions (level III) that can comprise multiple problem solving styles.

This approach succeeded in deriving adaptation-oriented knowledge discovering of new learners' behavior semantics from naturally occurring learning behavior. However, this multi-level clustering approach could not be applied to the complex behavioral models. Moreover, this approach should have used dynamic clustering techniques so as to be integrated into learning management systems.

Also, in [26], **Klasnja- Miliabcevic, A. et al.** proposed a recommendation module of Protus - a programming tutoring system - which can automatically adapt to the interests and knowledge levels of learners. It is based on clustering different learning styles, then analyzing the habits and the interests of the learners through mining the frequent sequences by the AprioriAll algorithm and finally, recommending the learning content according to the ratings of these frequent sequences. Although this system could improve the quality of an intelligent tutoring system, as well as keep the recommendation up-to-date, the recognition of learner's interests and knowledge level in different topics was not good and needed to be improved.

4. Comparison between the Semantic Personalization techniques

In this section, we will summarize the Semantic Personalization approaches discussed in section 3 as a table of comparison (**Table 1**). In this table, we are trying to analyze these approaches (techniques) through classifying them into categories (as shown in **Fig. 3**). Each category includes the techniques having similarities in their dependency, but are different – and thus can be compared - in the following points: the method of personalization, the suitable e-learning applications, the advantages and limitations.

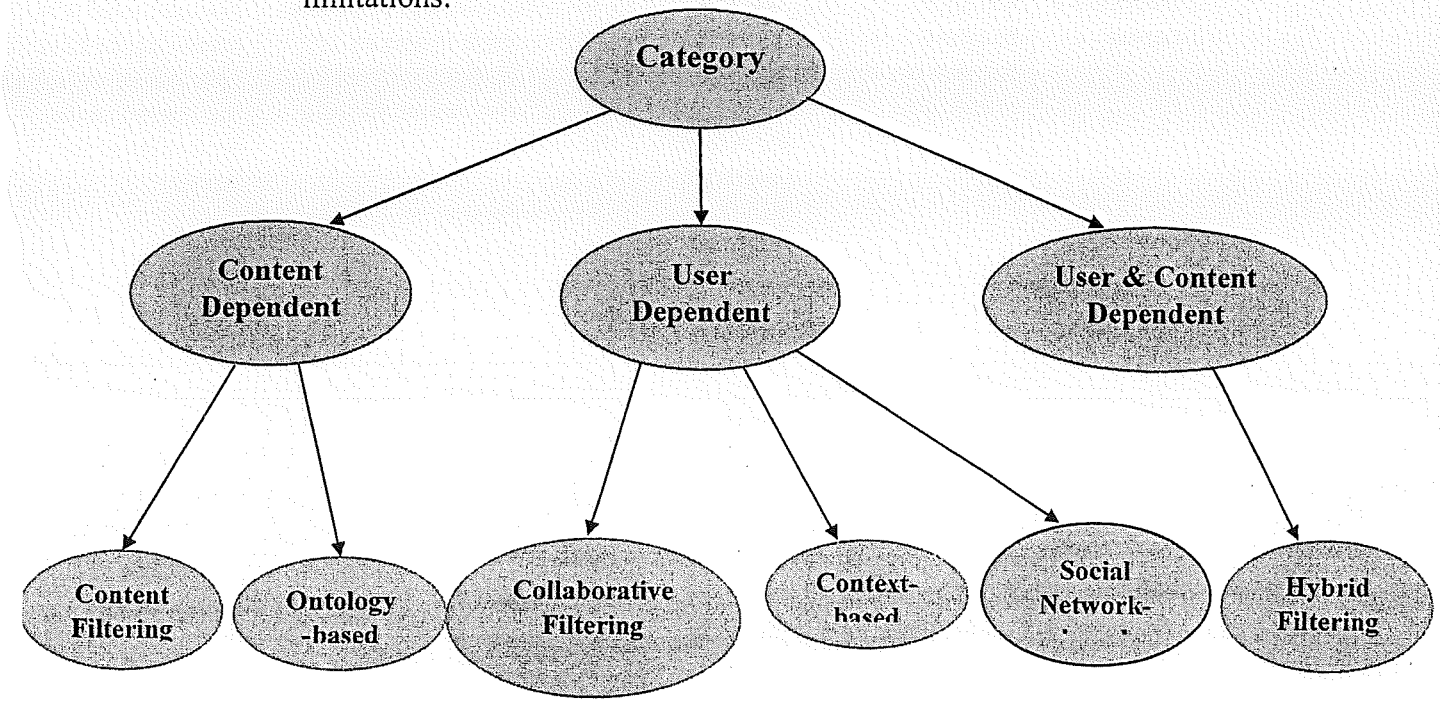


Fig. 3: "Categories of Semantic Personalization Approaches"

Table 1: "Comparison between the Semantic Personalization approaches"

Category	Approach	Method of Personalization	E-learning Application	Advantages	Limitations
Content Dependent	**Content Filtering**	Semantic Relations within one or across many items are computed	Suitable for online experienced learners, but want to improve their learning process through a learning platform that can be adapted to their needs.	• Combining the feedback from user with semantically similar items • Providing more interesting recommendations for concepts by using combinations of content feature and semantic relations	Not suitable for students who do not have online learning experience
	Ontology-Based	Classification of Domain Ontology	Suitable for the teachers taking benefits from the knowledge	Used to define restrictions on the instances that can be used by the user to reduce the total number of displayed instances	Just considers the domain ontology for personalization without considering the context information of the user which may vary with time.
User Dependent	**Collaborative Filtering**	Semantic Similarity between two users is computed	Adaptive Content, Adaptive Navigation and Adaptive Interface for each student.	• Mapping users and items to a set of concepts in domain ontology reducing the problem of insufficient resources (cold start problem) • Recommending items having high semantic similarity. Expressing User preferences as ratings	Not concerned with measuring the effect of prior learner experience and interest on the personality of the learner
	Context-Based	User context model and context aware computing	Suitable for academic students	• Context ontology sharing • Ontology reuse • Logic reasoning mechanisms	• Low Validity (tested only in academic situations but not tested in real situations) • Using Complex dynamic algorithm
	Social Network-based	Social Network and Semantic Analyses	• Information systems e-learning domain • Researchers	• Providing more interesting recommendations • Focusing on the social communication between the learners	Concerned only with the degree of interest not the context

Continuing Table 1: "Comparison between the Semantic Personalization approaches"

User & Content Dependent	**Hybrid Filtering**	Combining both the content and collaborative filtering techniques	Problem Solving E-learning domain	Takes the advantages of both filtering techniques to produce high quality recommendations.	More time has to be spent in designing the rules, sequences of activities and updating them when being applied on the analysis of web usage data.

From the above table, we can conclude that some of the discussed semantic e-learning personalization approaches are dependent on the content of the e-learning course while the others are dependent on the profile of the user (learner). So we can classify these approaches into two main categories: the first category is the content-dependent approaches including the content filtering and the ontology-based approaches. The approaches in this category are more suitable to the teachers rather than the students, but they may be used by very experienced learners. The second category – the user-dependent category - includes the collaborative filtering, the context-based and the social network-based approaches. The approaches in this category are more suitable to the learners and provide more adaptivity for different types of learners while the third category includes the hybrid filtering approach which combines both the user and the content dependencies, but it depends mainly on designing rules and sequences of activities which may consume more time than the other categories.

5. Conclusion

In this paper, we have overviewed six techniques commonly used for personalization of semantic web-based e-learning systems. These techniques are the: Ontology-based, Context-based, Social Network-based, Collaborative Filtering, Content Filtering and Hybrid Filtering techniques. We gathered the advantages and disadvantages of those techniques. Based on the above survey, the **Social Network-based** approach is more suitable for researchers, and it is capable for detecting the learner's way of learning through his/her profile, but it depends only on the degree of interest of the learner. The **Context-based** approach takes

the learners' context into consideration and thus makes the system more adaptable. However, it is so complex and dynamic and not suitable for undergraduate or school learners.

Since our research for personalization relies on adapting the learning materials to various categories of learners according to the type of the profile of each learner which can be detected through a social network model, thus our aim is to try to combine the Social Network-based and the Context-based approaches, taking the features of both approaches and integrating anyone of them with the Hybrid Filtering approach to obtain an interactive and a collaborative e-learning platform. This platform should helps users to build their personalized social network and knowledge network during the process of learning. It should also be capable of connecting a course network of users with their social network and knowledge network. Moreover, it should provide recommendation and search services to facilitate the interaction and the collaboration between the learners.

Acknowledgment

The authors would like to thank and appreciate **Dr. Eng. Fatma A. El-Licy**, a professor in the department of Computer & Information Sciences, Institute of Studies and Statistics Researches, Cairo University for her faithful assistance, support and potential help regarding this research.

6. References

[1] Lu, J.: A Personalized e-Learning Material Recommender System. Proceedings of the 2nd International Conference on Information Technology for Application (ICITA) (2004)

[2] Konstantinos, M., Penelope, M., Giannis, K. & Aglaia, L. T.: Semantic e-Learning, Next Generation of e-Learning. Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, pp. 295–308 (2010)

[3] Chrysafiadi, K. & Virvou, M.: Student modeling approaches, A literature review for the last decade. Expert Systems with Applications 40, pp. 4715–4729 (2013)

[4] Markowska-Kaczmar, U., Kwasnicka,H. & Paradowski, M.: Intelligent Techniques in Personalization of Learning in e-Learning Systems. Computational Intelligence for Tech. Enhanced Learning, SCI 273, pp. 1–23 (2010)

[5] Esichaikul, V., Lamnoi, S. & Bechter, C.: Student Modelling in Adaptive E-Learning Systems. Knowledge Management & E-Learning: An International Journal, Vol.3, No.3. (2011)

[6] Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American 284(5), 34–43 (2001)

[7] Vargas-Vera, M. & Lytras, M.: Personalized Learning Using Ontologies and Semantic Web Technologies. WSKS, LNAI 5288, pp. 177–186 (2008)

[8] Ameen, A., Khan, K. U. R. & Rani, B. P. Semantic Web Personalization: A Survey. Information and Knowledge Management, Vol. 2, No. 6 (2012)

[9] Othman, M. S., Suhaimi, S. M., Yusuf, L. M. , Yusof, N. and Mohamad, N.: An Analysis of Social Network Categories: Social Learning and Social Friendship. International Conference on Teaching and Learning in Higher Education (ICTLHE), Procedia - Social and Behavioral Sciences 56, pp. 441 – 447 (2012)

[10] Lichtnow, D., Gasparini, I., Bouzeghoub, A., De Oliveira, J. P. M. & Pimenta, M. S.: Recommendation of Learning Material through Students' Collaboration and User Modeling in an Adaptive E-Learning Environment. Technology-Enhanced Systems and Tools, SCI 350, pp. 257–278 (2011)

[11] Limongelli, C., Sciarrone, F. & Vaste , G.: Personalized e-learning in Moodle: the Moodle LS System. Journal of e-Learning and Knowledge Society (Je-LKS)- EN Vol. 7, n. 1, pp. 49 – 58 (2011)

[12] Gaeta, M., Miranda, S., Orciuoli, F., Paolozzi, S. & Poce, A.: An Approach To Personalized e-Learning. SYSTEMICS, CYBERNETICS AND INFORMATICS, VOLUME 11 – No. 1 (2013)

[13] Mencar, C., Castiello, C. & Fanelli, A. M.: A Profile Modelling Approach for E-Learning Systems. ICCSA, Part II, LNCS 5073, pp. 275–290 (2008)

[14] Eyharabide,V., Gasparini, I., Schiaffino, S., Pimenta,M. & Amandi,A.: Personalized e-Learning Environments: Considering Students' Contexts. International Federation for Information Processing (IFIP), AICT 302, pp. 48–57 (2009)

[15] Kurilovas, E., Zilinskiene, I., Dagiene,V.: Recommending suitable learning scenarios according to learners' preferences, An improved swarm based approach. Computers in Human Behavior 30, pp. 550–557 (2013)

[16] Jeremić, Z., Jovanović, J. & Gašević, D.: Personal Learning Environments on the Social Semantic Web. Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal (SWJ) (2011)

[17] Xu, Y., Guo, X., Hao, J., Ma, J., Lau, R. Y. K. & Xu, W.: Combining social network and semantic concept analysis for personalized academic researcher recommendation. Decision Support Systems 54, pp. 564–573 (2012)

[18] Mansur, A. B. F. & Yusof, N.: Social learning network analysis model to identify learning patterns using ontology clustering techniques and meaningful learning. Computers & Education 63, pp. 73–86 (2012)

[19] Salehi , M. , Kamalabadi, I. N. &Ghoushchi, M. B. G.: Personalized recommendation of learning material using sequential pattern mining and attribute based collaborative filtering. Springer Science+Business Media New York (2012)

[20] Kim, H. H.: A Personalized Recommendation Method Using a Tagging Ontology for a Social E-Learning System. ACIIDS , LNAI 6591, pp. 357–366 (2011)

[21] Klašnja-Milićević, A.: Personalized Recommendation Based on Collaborative Tagging Techniques for an E-learning System. Doctoral Dissertation, University of NOVISAD, Faculty of Science, Dept. of Mathematics & Informatics (2012)

[22] Parveen, R., Jaiswal, A. K. & Kant, V.: E-Learning Recommendation Systems – A Survey. International Journal of Engineering Research and Development, Volume 4, Issue 12, pp.10-12 (2012)

[23] Cheung, R. & Kazemian, H. B.: An Adaptive Framework for Personalized E-Learning. CCIS 136, pp. 292–306 (2011)

[24] Essalmi, F., Ayed, L. J. B., M., Jemni, M., Kinshuk & Graf, S.: A fully personalization strategy of E-learning scenarios. Computers in Human Behavior 26, pp. 581–591 (2010)

[25] Köck, M. & Paramythis, A.: Activity sequence modelling and dynamic clustering for personalized e-learning. Springer Science+Business Media B.V. (2011)

[26] Klasnja- Miliabcevic, A., Vesin, B., Ivanovic, M. & Budimac, Z.: E-Learning personalization based on hybrid recommendation strategy and learning style identification. Computers & Education 56, pp. 885–899 (2010)

Manipulating Roughness Uncertainty In Optimization Problems

Assem A. Alsawy[1] Hesham A. Hefny[2]

Abstract

The concept of rough set is an extension of the standard mathematical set. In this extension, an uncertain set is represented by its upper and lower approximations. In recent years some definitions for rough number were presented. In this work a new definition for rough number is proposed, and some types of rough numbers are introduced, also this work gives some suggested forms for the basic arithmetic operations on rough numbers, and finally, some examples are presented to illustrate the suggested definitions.

Keywords: Rough Set, Roughness, Rough Number, Rough Arithmetic, Optimization, Traveling Salesman Problems (TSP).

1. Introduction

In 1982, Pawlak proposed a concept of rough set as a new mathematical tool to deal with vagueness and uncertainty [1]. A relationship exists between fuzzy set theory and rough set theory [2]. Rough set theory does not compete with fuzzy set theory, with which it is frequently contrasted, but rather complements it. In any case, rough set theory and fuzzy set theory are independent approaches to imperfect knowledge. The problems using rough set theory include data reduction(i.e. elimination of superfluous data), discovery of data dependencies, estimation of data significance, generation of decision(control) algorithms from data, approximate classification of data, discovery of similarities or differences in data, discovery of patterns in data and discovery of cause effect relationships[3]. To define rough set formally [4], as shown in figure 1, let U be a set; let R be a set of equivalence relations imposed on the universe U.

[1] Dept. of Computer and Information Sciences Institute of Statistical Studies and Research Cairo University

[2] Prof. Head of Dept. of Computer and Information Sciences Institute of Statistical Studies and Research - Cairo University

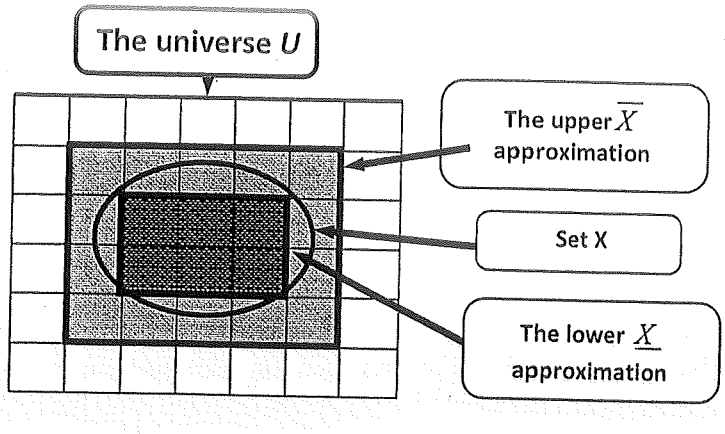


Figure 1: $X = (\underline{X}, \overline{X})$

The knowledge base K is defined as K= (U, R); and the concept X in K is defined as a subset of U, i.e., $X \subset U$. Further, the upper approximation of the concept X in K under a given equivalence relation R is defined as:

$$\overline{X} = \{x: [x]_R \cap X \neq 0\}.$$

And the lower approximation of the concept X is defined as:

$$\underline{X} = \{x: [x]_R \subset X\}$$

Here $[x]_R$ represents the equivalence class of x under the given relation R. If $\underline{X} = \overline{X}$, then the concept X is considered to be precise; otherwise, the concept is vague. For a vague concept, these two approximations under the given relation in the knowledge base K can be calculated and obtained. A rough set is a pair $(\underline{X}, \overline{X})$ of standard sets, the lower and upper approximation, the lower approximation $\underline{X}$, is always a subset of $\overline{X}$, the upper approximation. The meaning of these two sets is that if a data point lies in $\underline{X}$, we are sure that the point is in the rough set, if a data point lies in $X - \underline{X}$, we are unsure whether or not the point is in the rough set, and if a data point is outside $\overline{X}$, we are sure that the point is not in the rough set. These sets can contain either individual points, or continuous areas; As opposed to rough sets, standard sets are often called crisp, a term that also applies to a rough set with an empty area of uncertainty. Conversely, a rough set with an empty lower approximation and a non-empty area of uncertainty can be called completely rough. Here is an example for calculating $\underline{X}$ and $\overline{X}$:

Let $U = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ and the equivalence relation $R = \{Y_1, Y_2, Y_3, Y_4, Y_5\}$. Since $Y_1=\{x_1, x_5\}$, $Y_2=\{x_2\}$, $Y_3=\{x_3, x_4\}$, $Y_4=\{x_6\}$, $Y_5=\{x_7, x_8\}$

Two sets X_1 and X_2 such as: $X_1=\{x_1, x_3, x_7\}$, $X_2=\{x_2, x_4\}$

For Calculating $\underline{X}_1$, $\overline{X}_1$ and $\underline{X}_2$, $\overline{X}_2$ As $Y_1, Y_2, Y_3, Y_4, Y_5 \not\subset X_1$ hence $\underline{X}_1 = \Phi$ Also $Y_1 \cap X_1 \neq \Phi$ $Y_2 \cap X_1 \neq \Phi$ $Y_3 \cap X_1 \neq \Phi$ $Y_4 \cap X_1 = \Phi$ $Y_5 \cap X_1 = \Phi$ So $\overline{X}_1 = Y_1 \cup Y_3 \cup Y_5 = \{x_1, x_5, x_3, x_4, x_7, x_8\}$

Also, $\underline{X}_2 = \{x_2\}$, $\overline{X}_2 = \{x_2, x_3, x_4\}$.

2. Rough Number

In 1996, Pawlak defined a new class of real numbers called rough numbers [5]. Rough numbers are based on the several following notions. For the set R^+ a sequence of nonnegative real numbers $S = (x_i)_{i=1}^{n}$ such that $x_i < x_j$, $i < j$ is called the categorization of R^+. The approximation space is the ordered pair $A = (R^+, S)$. Every categorization S of R^+ induces partition $\pi(s)$ on R^+ defined as $\pi(S) = \{0,]0, x_1[, x_1, .., x_i,]x_i, x_{i+1}[, ..\}$ where $]x_i, x_{i+1}[$ denotes the open interval. $S(x)$ denotes the block of partition $\pi(S)$: $x \in S(x)$, as special case if $x \in S$, then $S(x) = \{x\}$

Let $x \in]x_i, x_{i+1}[$, the closed interval $\overline{S}(x) = [x_i, x_{i+1}]$ is called the closure $S(x)$.

$Q(x)$ is the closed interval $[0, x]$. For a given approximation space $A = (R^+, S)$ for every $Q(x)$ its lower and upper approximation may be defined, denoted as $\underline{S}(Q(x))$ and $\overline{S}(Q(x))$ respectively such that:

$$\underline{S}(Q(x)) = \{y \in R^+ : S(y) \subseteq Q(x)\} ;$$

$$\overline{S}(Q(x)) = \{y \in R^+ : S(y) \cap Q(x) \neq \varnothing\}.$$

As the S-lower and S-upper approximation of the nonnegative real number x the following values are considered:

$$\underline{S}(x) = Sup\{y \in S : y \leq x\};$$

$$\overline{S}(x) = Inf\{y \in S : y \geq x\}.$$

These notions lead to the definition of an approximation of a real number x on the basis of a set of real numbers $S : S(x) = (\underline{S}(x), \overline{S}(x))$. If $\underline{S}(x) = \overline{S}(x)$ then x *is* exact number, otherwise it is inexact (rough) number. Every inexact number may be represented as a pair of exact numbers. Here an example to illustrate this definition, for the inexact number $c \in] x_2, x_3 [$ as shown in figure 2, $Q(c)$ is the closed interval $[0, c]$. The lower approximation and upper approximation are:

$$\underline{S}(Q(c)) = \{y \in R^+ : S(y) \subseteq Q(c)\} = [0, x_2]$$

$$\overline{S}(Q(c)) = \{y \in R^+ : S(y) \cap Q(c) \neq \varnothing\} = [0, x_3]$$

Or $\underline{S}(c) = Sup\{y \in S : y \leq c\} = x_2$

$\overline{S}(c) = Inf\{y \in S : y \geq c\} = x_3$

The rough number c could be written as c= (x_2, x_3) as shown in figure 2.

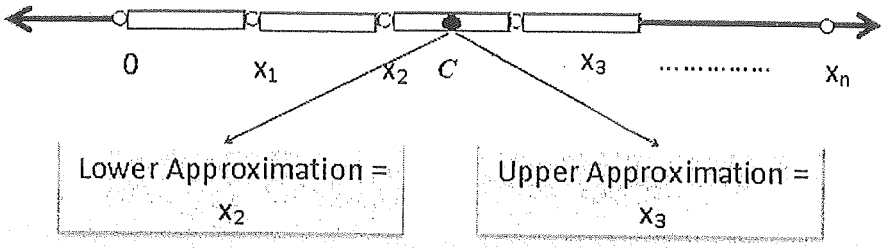


Figure2: Rough number c = (x_2, x_3)

The approximation space is $A = (R^+, N)$, N is a set of natural numbers

$\pi(N) = \{0,]0, 1[, 1,]1, 2[, 2,]2, 3[\ldots\}$

For the inexact number x: $4 < x < 5$ as shown in figure 3, the lower approximation and upper approximation are:

$\underline{N}([0, x]) = \{y \in R^+ : N(y) \subseteq [0, x]\} = [0, 4]$

$\overline{N}([0, x]) = \{y \in R^+ : N(y) \cap [0, x] \neq \varnothing\} = [0, 5]$

Or $\underline{N}(x) = Sup\{y \in N : y \leq x\} = 4$

$\overline{N}(x) = Inf\{y \in N : y \geq x\} = 5$

The rough number x could be written as $x = (4, 5)$ as shown in figure 3.

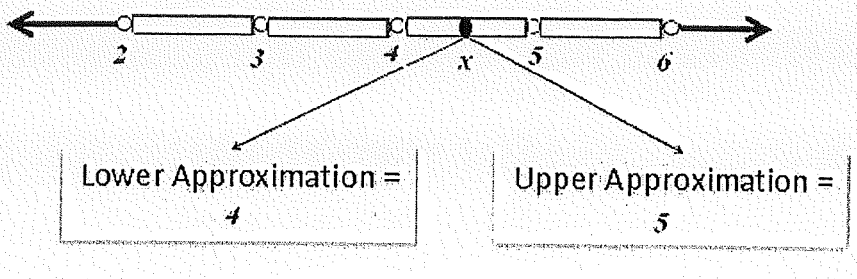


Figure 3: Rough number $x = (4, 5)$

3. Extension for rough number

In 2011 Marcin Michalak proposed an extension for the definitions of rough number as follows [6]: Let S is the categorization of R such that $S \subseteq R$. The approximation space is the ordered pair A = (R, S) for a given $S = (x_i)_{i=1}^{n}$ the generalized projection Π is defined:

$\pi(S) = \{]-\infty, x_1[, x_1,]x_1, x_2[, x_2, .., x_i,]x_i, x_{i+1}[, ..\}$ Notions $\underline{S}(x)$ and $\overline{S}(x)$ remain as Pawlak definition. $Q(x)$ should be defined as $Q(x) =]-\infty, x]$. This extension of the Pawlak definition leads us to the notion of rough real number. As an example the approximation space is $A = (R, Z)$, Z is a set of integer numbers

$\pi(Z) = \{..,]-2, -1[, -1,]-1, 0[, 0,]0, 1[, 1,]1, 2[, 2,]2, 3[, ..\}$

For the inexact number x: $-3 < x < -2$ the lower approximation and upper approximation are: $\underline{Z}(]-\infty, x]) = \{y \in R : Z(y) \subseteq]-\infty, x]\} =]-\infty, -3]$

$\overline{Z}(]-\infty, x]) = \{y \in R : Z(y) \cap]-\infty, x] \neq \varnothing\} =]-\infty, -2]$

Or $\underline{Z}(x) = Sup\{y \in Z : y \leq x\} = -3$

$\overline{Z}(x) = Inf\{y \in Z : y \geq x\} = -2$

The rough number x could be written as x = (-3, -2) as shown in figure 4.

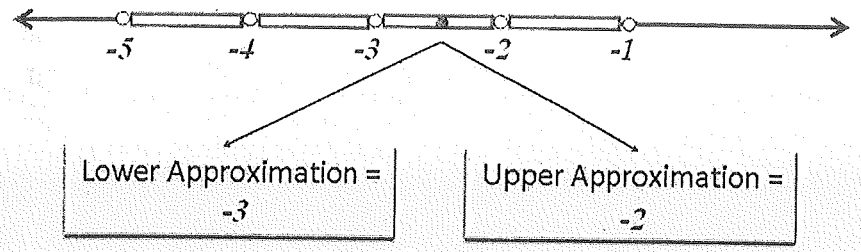


Figure 4: Rough number x = (-3, -2)

4. Proposed Definition for rough number

Definition: Rough number is a rough subset of real number line with two conditions:

1. Continuity
2. The lower approximation $\neq \varnothing$

For a set of real numbers R, a set of exact numbers $S = (x_i)_{i=1}^{n}$ are categorizations of R each categorization $s \in S$ induces partition of $\pi(S)$ defined as:

$\pi(S) = \{.....x_0,]x_0, x_1[, x_1, .., x_i,]x_i, x_{i+1}[, ..\}$ For any rough number X lower and upper approximation denoted as $\underline{X}$ and $\overline{X}$ respectively, since

$\underline{X} = \{y \in R : S(y) \subseteq X\}$, $\overline{X} = \{y \in R : S(y) \cap X \neq \varnothing\}$

Rough number X could be expressed by lower and upper approximation as: $X = (\underline{X}, \overline{X})$, rough number is a general form and from which a crisp number and interval are special cases

This proposed definition is a general form for rough number; somehow it includes the other ideas of the Pawlak and Michalak definitions

5. Types of rough numbers

According to new definition of rough number it can classify some types of rough numbers as follows:

(1) If $X = \{s_i\}$ and $s_i \in S$, as shown in figure 5, then $\underline{X} = \overline{X} = \{s_i\}$, that means X is exact number

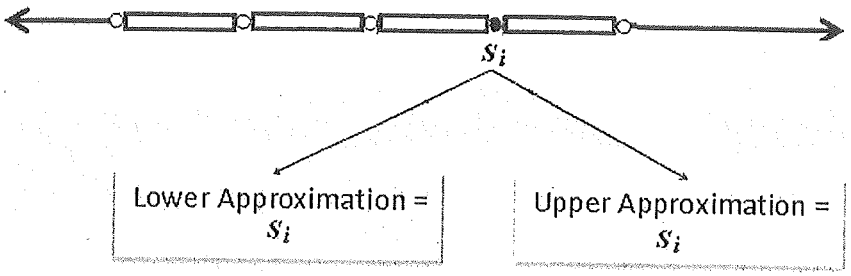


Figure 5: $X = \{s_i\}$

In rough number notation $X = (s_i, s_i)$ we call this type "exact single number"

(2) If $X = \{x_i\}$ and $x_i \notin S$, $x_i \in]\, s_j,\ s_{j+1}[$,as shown in figure 6, then $\underline{X} = \varnothing$, $\overline{X} = S(x_i) =]\, s_j,\ s_{j+1}[$

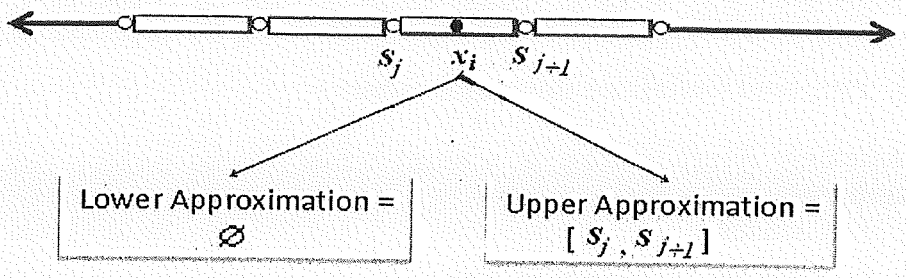


Figure 6: $X = \{x_i\}$ and $x_i \notin S$

In rough number notation $X = (\varnothing,]\, s_j,\ s_{j+1}[)$ we call this type "inexact single number".

(3) If $X = [\, s_i\ ,\ s_j]$, and $s_i, s_j \in S$ as shown in figure 7, then $\underline{X} = \overline{X} = [\, s_i\ ,\ s_j]$ that means X is exact interval

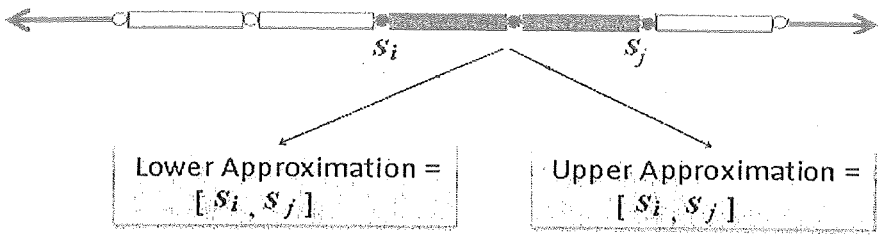


Figure 7: $X = [\, s_i , s_j]$

In rough number notation $X = ([s_j , s_{j+1}\,] , [s_j , s_{j+1}\,])$ we call this type "exact interval number".

(4) If $X = [x , y]$, x or $y \notin S$

- if a single exact number $s_i \in X$, as shown in figure 8, then $\underline{X} = s_i$, $\overline{X} = [\, s_{i-1} , s_{i+1}]$

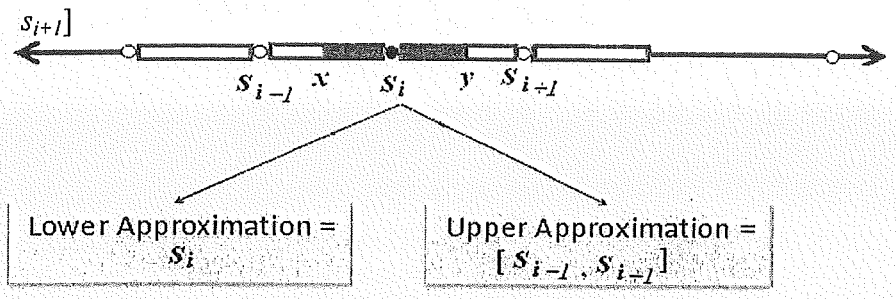


Figure 8: $X = [x , y]$

In rough number notation $X = (s_j , [s_{j-1} , s_{j+1}\,])$

- If exact interval $[\, s_i , s_j] \subset X$, as shown in figure 9, then $\underline{X} = [\, s_i , s_j]$, $\overline{X} = [\, s_{i-1} , s_{j+1}]$

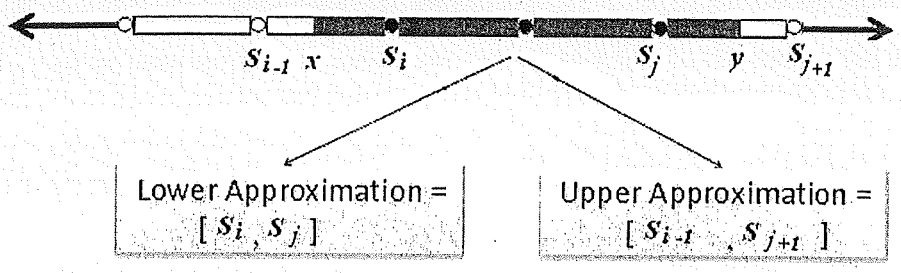


Figure 9: $X = [x , y]$

In rough number notation $X = ([\, s_i , s_j] , [s_{j-1} , s_{j+1}\,])$ we will call this type "inexact interval number".

5.1 Numerical examples

If $\square = \{..., -2, -1, 0, 1, 2, 3, ...\}$ is the set of all integer numbers are categorizations for $\square$ the set of real numbers, every categorization in $\square$ induces a partition in $\pi(\square)$ which defined as follows:

$\pi(\square) = \{ . .,]-2, -1[, -1,]-1, 0[, 0,]0, 1[, 1,]1, 2[, 2,]2, 3[, . .\}$

Here, some examples to approximate some values using the concepts of rough numbers are given as follows:

Example (1): for inexact single number X= $\{x_i\}$, $x_i \in]2, 3[$
Hence $\underline{X} = \varnothing$, $\overline{X} =]2, 3[$ therefore X= $(\varnothing,]2, 3[)$

Example (2): for inexact interval number X= $[x_i , x_j]$, $x_i \in]-1, 0[$, $x_j \in]0, 1[$
Hence $\underline{X} = \{0\}$, $\overline{X} =]-1, 1[$ therefore X= $(0,]-1, 1[)$

Example (3): for inexact interval number X= $[x_i , x_j]$, $x_i \in]1, 2[$, $x_j \in]3, 4[$
Hence $\underline{X} = [2, 3]$, $\overline{X} =]1, 4[$ therefore X= $([2, 3],]1, 4[)$

Example (4): for inexact interval number X= $[x_i , x_j]$, $x_i = 1$, $x_j \in]4, 5[$
Hence $\underline{X} = [1, 4]$, $\overline{X} = [1, 5[$ therefore X= $([1, 4], [1, 5[)$
The proposed formula for rough number allows us to represent the linguistic values, as example, the value "around 5" could be expressed as rough number as shown in figure 10 and it could be written in rough notation as: around 5=(5,]4,6[ )

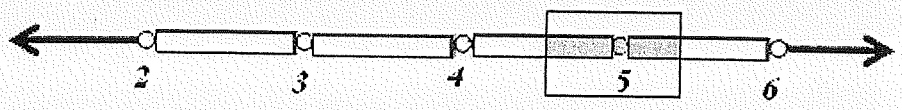


Figure 10: Around 5

Also linguistic variable as temperature could be represented by using rough number notation, as example, if temperature has values (cold, cool, hot) as shown in figure 11, these values could be considered as rough number as follows:
Cold = $(]-\infty, 10],]-\infty, 15[)$, Cool = $([15, 25],]10, 30[)$, Hot = $([30, \infty[,]25, \infty[)$.

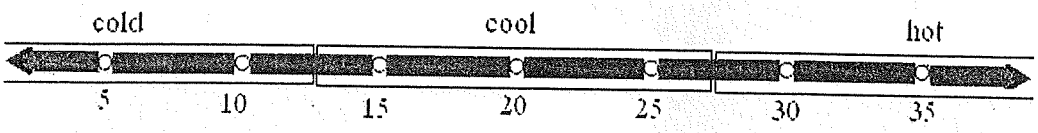


Figure 11: Temperature values

6. Arithmetic on rough numbers

The greatest benefit of using the concept of rough number rather than the concept of rough set is the ability to define and perform arithmetic operations. Since the concept of rough number depends mainly on the intervals, this section gives some review of arithmetic on intervals and then some basic operations on rough numbers will be introduced

6.1 Definition [7]: Let X and Y are real intervals, then their sum, difference and product are defined to be the following sets:

$X + Y = \{x + y \mid x \in X \wedge y \in Y\}$

$X - Y = \{x - y \mid x \in X \wedge y \in Y\}$

$X^* Y = \{x * y \mid x \in X \wedge y \in Y\}$

$X / Y = \{x / y \mid x \in X \wedge y \in Y, y \neq 0\}$

According to this definition, some theories are coined to simplify the arithmetic basic operations as follows [8]: For two intervals [a, b] and [c, d] which are subsets of the real line$]-\infty, \infty[$

$[a, b] + [c, d] = [\min (a + c, a + d, b + c, b + d), \max (a + c, a + d, b + c, b + d)]$
$= [a + c, b + d].$

$[a, b] - [c, d] = [\min (a - c, a - d, b - c, b - d), \max (a - c, a - d, b - c, b - d)]$
$= [a - d, b - c].$

$[a, b] \times [c, d] = [\min (a \times c, a \times d, b \times c, b \times d), \max (a \times c, a \times d, b \times c, b \times d)]$

$[a, b] / [c, d] = [\min (a / c, a / d, b / c, b /d), \max (a /c, a /d, b /c, b /d)]$ when 0 is not in [c, d].

The following subsections introduce the suggested formulas for basic operations on rough numbers

6.2 Addition

For two rough numbers $X = (\underline{X}, \overline{X})$ and $Y = (\underline{Y}, \overline{Y})$ addition operation could be defined as: $X+Y = (\underline{X}+\underline{Y}, \overline{X}+\overline{Y})$, that means X+Y is a rough number with lower and upper approximations as follows: $\underline{X+Y} = \underline{X}+\underline{Y}$ and $\overline{X+Y} = \overline{X}+\overline{Y}$.

Example (5): for two rough numbers X= (2,] 1, 3[), and Y = (4, [3, 5])
X+Y= (6,] 4, 8[)

Example (6): for two rough numbers X= ([-1, 0], [-2, 2]), and Y = (2, [2, 3])
X+Y= ([1, 2], [0, 5])

6.3 Subtraction

For two rough numbers $X = (\underline{X}, \overline{X})$ and $Y = (\underline{Y}, \overline{Y})$ subtraction operation could be defined as: $X-Y = (\underline{X}-\underline{Y}, \overline{X}-\overline{Y})$ that means X – Y is a rough number

with lower and upper approximations as follows: $\underline{X-Y}=\underline{X}-\underline{Y}$ and $\overline{X-Y}=\overline{X}-\overline{Y}$.

Example (7): for two rough numbers X= (2,] 1, 3[), and Y = (4,] 3, 5[)
X – Y= (-2,] – 4, 0[)
Example (8): for two rough numbers X= ([-1, 0],]-2, 2[), and Y = (2,] 2, 3[)
Then X- Y= ([-3, -2],]-5, 0[)

6.4 Multiplication

For two rough numbers X = ($\underline{X}$, $\overline{X}$) and Y= ($\underline{Y}$, $\overline{Y}$) multiplication operation could be defined as: X*Y or XY = ($\underline{X}*\underline{Y}$, $\overline{X}*\overline{Y}$) that means X * Y is a rough number with lower and upper approximations as follows: $\underline{X*Y}=\underline{X}*\underline{Y}$ and $\overline{X*Y}=\overline{X}*\overline{Y}$.

Example (9): for two rough numbers X= (2,] 1, 3[), and Y = (4,] 3, 5[)
Then X * Y= (8,] 3, 15[)
Example (10): for two rough numbers X= ([-1, 0], [-2, 2]), and Y = (2, [2, 3])
Then X * Y= ([-2, 0], [- 4, 6])

6.5 Division

For two rough numbers X = ($\underline{X}$,$\overline{X}$) and Y= ($\underline{Y}$,$\overline{Y}$) division operation could be defined as: X / Y = ($\underline{X/Y}$,$\overline{X/Y}$) that means X / Y is a rough number with lower and upper approximations as follows: $\underline{X/Y}=\underline{X}/\underline{Y}$ and $\overline{X/Y}=\overline{X}/\overline{Y}$.

Example (11): for two rough numbers X= (2,] 1, 3[), and Y = (4, [3, 5])
Then $\underline{X}/\underline{Y}$ = 0.5 therefore $\underline{X/Y}=\underline{X}/\underline{Y}=\varnothing$

$\overline{X}/\overline{Y}$ =] 1/5, 1 [, therefore $\overline{X/Y}=\overline{X}/\overline{Y}$ =] 0, 1[
X / Y = (∅,] 0, 1[]
Example (12): for two rough numbers X= ([-1, 0], [-2, 2]), and Y = (2, [2, 3])
Then $\underline{X}/\underline{Y}$ = [-0.5, 0] therefore $\underline{X/Y}=\underline{X}/\underline{Y}=\{0\}$

$\overline{X}/\overline{Y}$ =] -2/3, 1 [, therefore $\overline{X/Y}=\overline{X}/\overline{Y}$ =] -1, 1[
X / Y = (0,] -1, 1[]

5.1 Uncertain Travelling Salesman Problem (UTSP)

Travelling Salesman Problem (TSP) is widely known problem from class (NP) [9], [10]. In this problem, the salesman has to visit all the cities and return back to home at the end of journey [11], uncertain traveling salesman problem (UTSP) is TSP with uncertain costs (distances) [12], some solution for TSP in fuzzy

environment have been introduced [13], in next example shows the solution of TSP in rough environment.

Example: This example shows how to manipulate roughness uncertainty in uncertain traveling salesman problem. Given a list of cities and their pairwise distances, which are rough numbers, the set of cities is C = {C_1, C_2, C_3, C_4} as shown in figure 12.

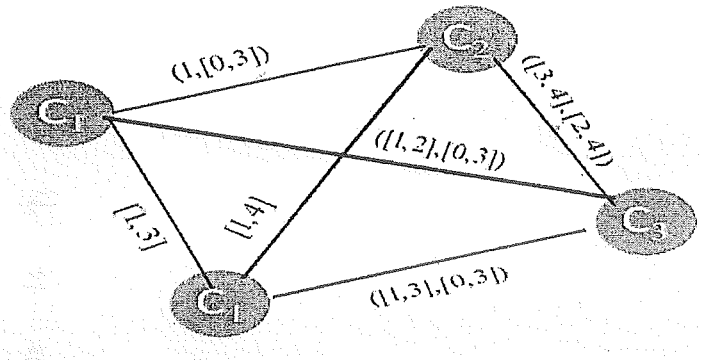


Figure 12: Rough Based Traveling Salesman Problem (RTSP)

D: is the distance function, it can be defined as following:

$D(C_1,C_2) = (1,[0,3])$

$D(C_1,C_3) = ([1,2],[0,3])$

$D(C_1,C_4) = [1,3]$

$D(C_2,C_3) = ([3,4],[2,4])$

$D(C_2,C_4) = [1,4]$

$D(C_3,C_4) = ([1,3],[0,3])$

To find the shortest cycle, first by calculating the lengths of all possible paths then ranking them according to the value of center value of upper approximation, we can say the shortest cycle is:

$$C_1 \rightarrow C_2 \rightarrow C_4 \rightarrow C_3 \rightarrow C_1$$

The length of this cycle can be calculated as follows:

$$(1,[0,3]) + [1,4] + ([1,3],[0,3]) + ([1,2],[0,3]) = ([4,10],[1,13])$$

Center value of upper approximation could be considered as a single value which reflects previous total distance which is: $\frac{1}{2}(1+13) = 7$, which is the same result in case of crisp number.

7. Conclusion

The proposed definition is a general formula for rough number; it includes the other ideas for the other definitions. The new definition allows us to represent the linguistic values as rough numbers. The greatest benefit of using the concept of rough number rather than the concept of rough set is the ability to define and perform arithmetic operations. New definitions of rough arithmetic enable us to solve optimization problems in rough environment.

8. References

[1] Pawlak. Z, "Rough Sets", International Journal of Computer and Information Science, Vol. 11, pp.341-356 , (1982)

[2] Pawlak. Z and Slowinski. R, "Rough Set approach to multiattribute decision analysis", Invited Review, Eur. Journal of Oper. Res., Vol.72, pp. 443-459, (1994)

[3] Komorouski.J, Pawlak.Z, Polkowski.L,Skowron.A, "Rough Sets: A tutorial",In Pal.S.K,Skowron.A (Eds) Rough Fuzzy Hybridization: A new trend in Decision making, springer, singapore, pp.3-98 , (1999)

[4] Szladow. A and Ziarko. W, " Rough sets: Working with imperfect data", AI Expert, 7, , pp. 36-41 (1993).

[5] Pawlak, Z.: Rough sets, rough relations and rough functions. Fundamental Informaticae 27(2-3), 103–108 (1996)

[6] Marcin Michalak "Rough Numbers and Rough Regression Rough Sets", Lecture Notes in Computer Science Volume 6743, pp 68-71, (2011)

[7] Hanson, R.J.. Interval arithmetic as a closed arithmetic system on a computer. Technical Report 197, Jet Propulsion Laboratory, (1968).

[8] Timothy J. Hickey, "Analytic Constraint Solving and Interval Arithmetic" Proceedings of the 27th Annual ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages, (2000).

[9] Applegate, David L., Robert E. Bixby, Vasek Chvatal, and William J. Cook. The traveling salesman problem: a computational study. Princeton University Press, (2011).

[10] Gutin, Gregory, and Abraham P. Punnen, eds. The traveling salesman problem and its variations. Vol. 12. Springer, (2002).

[11] Chvátal, Vašek, William Cook, George B. Dantzig, Delbert R. Fulkerson, and Selmer M. Johnson. "Solution of a Large-Scale Traveling-Salesman Problem." In50 Years of Integer Programming 1958-2008, pp. 7-28. Springer Berlin Heidelberg, (2010).

[12] Zhang, Bo, and Jin Peng. "Uncertain traveling salesman problem.", available at: http://www.orsc.edu.cn/online/110731.pdf , last visit: march 2011, (2011).

[13] Alsawy, Assem Ahmed, Hesham Ahmed Hefny, and Fatma El_Licy. "Fuzzy ant colony optimization algorithm." In Informatics and Systems (INFOS), 2010 The 7th International Conference on, pp. 1-5. IEEE, (2010).





جامعة القاهرة

معهد الدراسات والبحوث الإحصائية

المؤتمر السنوى الثامن والأربعون للإحصاء وعلوم الحاسب وبحوث العمليات

مجلـــــــد
علوم حاسب

٢٣-٢٦ ديسمبر ٢٠١٣