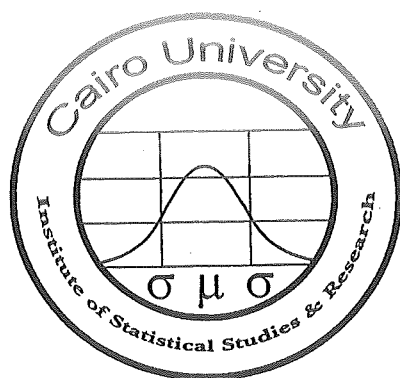




Cairo University
Institute of Statistical Studies and Research

The 47th Annual
Conference
on Statistics,
Computer Sciences,
and Operation Research

Computer Sciences

24-27, Dec . 2012

Handwritten text, mostly illegible due to fading and bleed-through. The text appears to be organized into several paragraphs, with some lines being more distinct than others. The handwriting is cursive and somewhat slanted.

Handwritten text, possibly a signature or a specific heading, located below the main body of text.

Handwritten text at the bottom of the page, continuing the narrative or providing a conclusion. The text is less dense than the upper sections.

Index

Computer Sciences

1	Computational Arabic Morphological Analyzer and Generator Systems: Theoretical Survey Maabid A , Elghazaly T.	1- 15
2	Enhancing UML Conceptual Modeling Through The Use of Integration Farouk N , Salah A , Kamel A	16-23
3	Human like Query Tool: Non-Stationary Fuzzy Set Approach Sabour.A.A , Rahman.B.M , Tawab.A.H , Gadallah.A.M , Hefny.H.A	24-37
4	Access Control Models for Vector- based Data Mennatallah H. Ibrahim , Nermin Hamza , Mohamed Nour El-Deen , Hesham A. Hefny	38-48
5	An Enhanced Model for Linguistic-based Fuzzy Ontology Zeinab E. Alarab , Ahmed M. Gadallah , Hesham A. Hefny	49-62
6	Granular Computing Approach for Manipulating Fuzziness Uncertainty Assem Ahmed Alsawy , Hesham A. Hefny	63-81
7	An Enhanced Fuzzy Query Approach for Relational Databases Sabour.A.A. , Gadallah. A.M , Hefny.H.A	82-96
8	A Survey of Extracting Arabic Multiword Expressions Samah Meghawry , Tatek Elghazaly , Abeer El-Korany , Akram Salah	97-109
9	A Gaussian Fuzzy ANP Model to Rank Different Scenarios Electrical Power Generation in Egypt Hesham A. Hefny, H.M.Elsayed , H.F.Aly	110-128
10	Reverse Engineering of PHP Grammar Mohamed M.S , El-Licy F.A , Salah A.I	129-146
11	Hand Gesture Recognition-Based Generative Versus Discriminative Models Elmezain.M, El-Shinawy.S	147-159
12	Arabic Named Entity Recognition for a New Corpus of Egyptian People 's Assembly Ahmed Montasser H.Farag	160-169
13	Applied Data Mining Classification Techniques for Diagnosing Breast Cancer Grades in Egypt Rabab Emad Mohammed	170-183
14	Improving the Behavior of QPSO Using Various Contraction Coefficient Forms Tarek M.Elbarbary , Hesham A.Hefny , Atef M.Abel Moneim	184-198
15	Analysis of Advanced Encryption Standard Algorithm with Constant and Dynamic Rotation I.A.Ismail, Galal H.Galal-Eldeen, Sherif Khattab, Mohamed Abd Elhamid	199-215

Computational Arabic Morphological Analyzer and Generator Systems: Theoretical Survey

Maabid A. , Elghazaly T.

Abstract

Morphology analysis is an essential part of most applications of natural language processing (NLP) which included different applications like Machine Translation (MT) and language rule based Information Retrieval (IR). Many Arabic morphological systems had built for different purposes with different algorithms and approaches; this paper is considered a survey of Arabic Morphological system from researchers' perspectives and approaches used to build them. Based on this survey; in the first part of this paper; the perspective of Arabic morphological systems had been classified into two major issues; one of them is the theoretical perspective and the second is the computational perspective of Arabic morphology. While the second part of this paper deals with approaches used to build the Arabic Morphology systems itself which are Table Lookup Approach, Combinatorial Approach, Linguistic Approach, Traditional Approaches, Finite-state Automata and Two-Level Morphology Approach and Pattern-Based Approach.

Keywords- Morphology, Arabic Morphology, NLP, Computational Morphology, Morphology Approaches

1. Introduction

Arabic is one of the languages having the characteristics that from one root the derivational and inflectional systems are able to produce a large number of words (lexical forms) each having specific patterns and semantics. Morphology in linguistics concerns with the study of the structure of words. In other words, morphology is simply a term for that branch of linguistics concerned with the forms words take in their different uses and construction. Morphology analysis is an essential part of most applications of natural language processing, NLP. Applications of NLP include Machine Translation, MT., Information Retrieval, IR., etc... While processing natural language applications are based on the Morphology part, it has to meet some criteria in order to satisfy the required functionality.

In Arabic, there are about 10,000 roots and 900 patterns. Not every root is compatible with every pattern. Linguistic experts must be consulted when creating comprehensive compatibility vectors. If compatibility vectors for all roots are available, then the general linguistic algorithm can be used. Once

found, roots should be checked in the vectors to verify the compatibility between the roots and the suggested patterns.

For noun analysis, a large number of lists need to be investigated, such as noun derivatives (e.g., nouns of place, of time, of instrument), nonstandard or solid nouns and different forms (e.g., plural, dual, and feminine). The authors state that the system covers the Arabic dictionary and can analyze vowelized and non-vowelized Arabic words. They also state that their implementation of pattern matching requires almost no time and lists do not require vast memory space.

Whereas many Arabic morphological systems are built with different strategies and methodologies this paper is considered a theoretical survey of Arabic morphological systems in two parts; the first part deals with the perspectives of Arabic Morphologies. The second part deals with the different approaches used in some Arabic Morphologies.

2. Perspectives of Arabic Morphology

2.1. Theoretical Perspective of Arabic Morphology

The notion of the morpheme as a meaningful string of segments delimited by the morpheme boundary mainly considered as blank spaces and sentences end marks, and containing no internal morpheme boundary, is challenged by the facts of Arabic Morphology, which exhibits properties that can be recalled [1], in very short words, as follows:

- Roots: built of three or four consonants. Each root dominates a clustering of Arabic lexical morphemes around a semantic field, which can be single, subdivided or multiple.
- Certain changes in nouns, verbs or adjectives based on these consonantal roots yield derivatives. Some vowel and syllabic patterns seem, subsequently, to be associated with a constant set of meanings.
- Traditional treatment of Arabic morphology – especially in computational morphology – sometimes remains taxonomic, abstracting away from the particular root and citing or generating all possible patterns [2].

2.2. Computational Perspective of Arabic Morphology

2.2.1 Knowledge-based Approaches

The computational treatments of Arabic morphology have been classified into the following major linguistic frameworks:

A. Syllable-based Morphology (SBM)

European languages especially Germanic language dealt with their morphology in terms of their syllable structure of word, thus this concept has been intended to apply to all languages [3]. Semitic languages especially Arabic language appears on the surface to offer biggest challenge to this theory [4]. Syllable strings therefore have the final syllable as syl1, the penultimate syllable as syl2 and so on. This ignores higher level structures such as feet or tone groups, as well as lower level structures such as mora. In Arabic it must therefore define a default syllable structure for trilateral stems, with default positions for the stem consonants; assuming that the simplest structure as exemplified by the stem "KATAB". The standard phonotactics of Arabic require syllables of CV (preferred) or CVC structure. The root must be divided into two syllables, because there are three consonants, and syllables in Arabic may have a maximum of two consonants

B. Root-and-Pattern Morphology

McCarthy proposed a claim that the stems are formed by a derivational combination of a root morpheme and a vowel melody. Roots are said to overlap with patterns to form stems. McCarthy's analysis differs from Harris' [5] in abstracting out or auto segmentation of the vowels from the pattern and placing them on a separate tier of the analysis [6, 7].

McCarthy's auto segmental approach is reflected in most of the computational attempts to model Arabic morphology, especially in the systems written within finite state morphology [8, 9]. It would be useful to briefly review one of the largest systems ever built for Arabic morphology on the basis of finite-state technology, namely the Arabic morphology system implemented using Xerox finite state technology. Verbs in Arabic are either trilateral (having three radical letters) as "ktb" (breaking) or quadrilateral (having four) as "dHrj" (rolling), for the root "ktb" if the pattern "_a_a_" applied then a new stem "katab" can be produced.

C. Lexeme-based Morphology

Lexeme-base morphology supports the claim that the stem is the only morphologically relevant form of a lexeme [10]. Lexeme-based morphology is (usually) an Item-and-Process approach. Instead of analyzing a word form as a set of morphemes arranged in sequence, a word form is said to be the result of applying rules that alter a word form or stem in order to produce a new one. An inflectional rule takes a stem, changes it as is required by the rule, and outputs a word form; a derivational rule takes a stem, changes it as per its own

requirements, and outputs a derived stem; a compounding rule takes word forms, and similarly outputs a compound stem. In lexeme-based Morphology; given the forms 'cat' and 'cats', we would say that there is a lexeme [cat] that has two word forms 'cat' and 'cats'

D. Morpheme-based morphology

In morpheme-based morphology, word forms are analyzed as arrangements of morphemes. A morpheme is defined as the minimal meaningful unit of a language. In a word like independently, the morphemes are in-, depend, -ent, and ly; depend is the root and the other morphemes are, in this case, derivational affixes. In a word like dogs, we say that dog is the root, and that -s is an inflectional morpheme. In its simplest (and most naïve) form, this way of analyzing word forms treats words as if they were made of morphemes put after each other like beads on a string, is called Item-and-Arrangement. More modern and sophisticated approaches seek to maintain the idea of the morpheme while accommodating non-concatenative, analogical, and other processes that have proven problematic for Item-and-Arrangement theories and similar approaches[11]. Morpheme-based morphology presumes three basic:

- Baudoin's SINGLE MORPHEME HYPOTHESIS: Roots and affixes have the same status in the theory, they are MORPHEMES.
- Bloomfield's SIGN BASE MORPHEME HYPOTHESIS: As morphemes, they are dualistic signs, since they have both (phonological) form and meaning.
- Bloomfield's LEXICAL MORPHEME HYPOTHESIS: The morphemes, affixes and roots alike, are stored in the lexicon.

Morpheme-based morphology comes in two flavors, one Bloom fieldian and one Hockettian. For Bloomfield, the morpheme was the minimal form with meaning, but it was not meaning itself. For Hockett, morphemes are meaning elements, not form elements. For him, there is a morpheme plural, with the allomorphs -s, -en, -ren etc. Within much morpheme-based morphological theory, these two views are mixed in unsystematic ways, so that a writer may talk about "the morpheme plural" and "the morpheme -s" in the same sentence, although these are different things [12].

E. Word-based morphology

Word-based morphology is (usually) a Word-and-paradigm approach. This theory takes paradigms as a central notion. Instead of stating rules to combine morphemes into word forms, or to generate word forms from stems, word-based morphology states generalizations that hold between the forms of

inflectional paradigms. The major point behind this approach is that many such generalizations are hard to state with either of the other approaches. The examples are usually drawn from fissional languages, where a given "piece" of a word, which a morpheme-based theory would call an inflectional morpheme, corresponds to a combination of grammatical categories, for example, "third person plural." Morpheme-based theories usually have no problems with this situation, since one just says that a given morpheme has two categories. Item-and-Process theories, on the other hand, often break down in cases like these, because they all too often assume that there will be two separate rules here, one for third person, and the other for plural, but the distinction between them turns out to be artificial. Word-and-Paradigm approaches treat these as whole words that are related to each other by analogical rules. Words can be categorized based on the pattern they fit into. This applies both to existing words and to new ones. Application of a pattern different from the one that has been used historically can give rise to a new word, such as older replacing elder (where older follows the normal pattern of adjectival superlatives) and cows replacing kine (where cows fits the regular pattern of plural formation).

F. Stem-based Lexicon

The central claim of this approach is that stem dependent on lexical databases, with entries associated with grammar and lexis specifications, is the most appropriate organization for the storage of pertinent information for Arabic.

Also known as "Stem-based Lexical Resources, Including Root-Pattern and Grammar-lexis Information"

Table (1) Knowledge-based Approaches

Approach	Claim	Use of word	Use of Stem	Use of Morpheme
Syllable-based Morphology	Deal with morphology in terms of their syllable structure of word	Yes	No	Yes
Root-and-Pattern Morphology	stems are formed by a derivational combination of a root morpheme and a vowel	No	Yes	Yes
Lexeme-based Morphology	The stem is the only morphologically relevant form of a lexeme	No	Yes	No
Morpheme-based morphology	Word forms are analyzed as arrangements of morphemes	Yes	No	Yes
Word-based morphology	states generalizations that hold between the forms of inflectional paradigms	Yes	No	Yes
Stem-based Lexicon	Stem dependent on lexical databases, with entries associated with grammar and lexis specifications	Yes	Yes	Yes

2.2.2 Empirical Approaches

A major benefit of the knowledge-based methods is that the rules and constraints for recognizing and classifying the internal structure of words are defined on a precise linguistic basis. Thus, under the assumption that the set of morphological rules and constraints define a linguistically consistent system, then, if a word can be morphologically analyzed, we can be sure that the resulting structure is correct [13]. Of course, this requires the computational basis of the analysis to be sound and complete; however, since we also assume this for the computational basis of the empirical methods, we do not consider this as a unique feature of knowledge-based methods. Furthermore, it is also often assumed that the modeled linguistic system is domain independent that is valid and applicable in any domain. As a consequence, the ultimate goal is to implement a linguistic knowledge base -in our case, a morphological system- that covers all possible allowable structures and only these. However, this requires not only that all possible allowable structures are known by the linguist, but that they can be formalized and implemented consistently preferable as non-redundant as possible [14].

In large scale implementation of Arabic NLP components the following drawbacks many raised:

- **Ambiguity:** The aim of formalizing all possible allowable structures means that an NLP component is to be expected to return all possible analyses (or readings) for a given input for further processing (by a human or another NLP component), as long as the system does not dispose of any decision criteria on how to rank or select between the alternative readings relative to a given domain or application.
- **Coverage:** Even if a linguistic domain is completely understood, it is extremely challenging to provide a complete implementation of all phenomena from scratch, because it might still be unclear how to represent a certain linguistic entity using the implementation formalism at hand or because not all possible ways and constraints (e.g., about the nature of the input data) are known in order to properly embed the NLP component into a larger application context.

As a solution direction for these kinds of problems, empirical-based methods are explored and developed in computational linguistics since the 1990s [15]. Empirical methods employ machine learning techniques to automatically extract linguistic knowledge from natural language data directly rather than require the system developer to manually encode the requisite knowledge. Since these methods are by definition data-driven, they actually also learn how

to weight between alternative solutions and how to predict useful information (e.g., a missing class label) for unknown entities through a rigorous statistically analysis of the data. In the beginning of the development of the new field of empirical methods for NLP, the proposed approaches have often been considered as alternative or even competing approaches to the corresponding knowledge-based methods [16]. However, in recent years the trend has become to consider both approaches more as complementary to each other, and new ways of integrating knowledge-based and empirical-methods are envisaged and actively explored.

3. Arabic Morphology Approaches

Many Arabic stemming algorithms have been developed, where each one has been implemented according specific methodology, in the subsequent sections we will represent an overview of such stemming algorithms used for Arabic systems.

3.1 Table Lookup Approach

The table lookup approach depends mainly on very large tables listed in alphabetical order storing Arabic words found in natural texts with their corresponding morphological parts. Such parts include stem, root, and affixation. Words may include functional words, foreign words, and proper names where each word uses a unique entry in the table. Multiple entries may exist for words spelled the same to reflect the possibility of multiple analyses [17].

3.2. Combinatorial Approach

Combinatorial approaches compare tested words against prepared lists for roots, patterns, particles, and affixes. The comparison is based on a combinatorial algorithm that tests all combinations of three or four letters of a given word in order to extract the root. In general, such approaches are simple but take a long time to process and require very large lists.

Al-Fedaghi and Al-Anzi [18] enhanced an algorithm proposed previously by Al-Fedaghi and Al-Sadoun. The enhancement is achieved by solving some important phonological and orthographical problems, such as assimilation, mutation, vocalization, and germination. The modified algorithm uses lists of trilateral roots and patterns with all combinations of affixes. It starts by comparing the input word against the pattern list to extract the root. The algorithm then runs in four modes to cover all possible orthographical and phonological cases.

3.3. Linguistic Approach

Linguistic-based approaches require a deep analysis of the Arabic morphological system. They simulate the behavior of a linguist during the analysis of a given Arabic word. In this approach, tested words are compared against lists of affixes or function words to derive stems and then compared against lists of patterns and roots in order to derive roots.

In general, such linguistic approaches are more accurate but require too many lists that need to be prepared and checked linguistically. Such lists incur a time overhead for accessing them. In addition, the mechanism for removing affixes is almost a process of trial and error, in which results are not guaranteed to be accurate.

Linguistic approaches can be subdivided into **traditional** or **cut-and-paste** approaches, and **finite-state automata** and **two-level morphology** approaches, and **Pattern-Based** Approach. In the subsequent section, traditional linguistic approaches will be discussed, followed by finite-state and two-level morphology models.

3.3.1 Traditional Approaches

In this section, a survey of Arabic morphological analysis techniques is conducted that includes a description of the algorithm, the method of its implementation.

Al-Khalil Arabic Morphological System is java compiled trial version applications published on April 2010. The system can analyze word or sentences typed in the text area [19]. The system can analyzes up to 10 words per second in small text up to 35 words per second in larger documents.

Al-Bawab [20, 21] proposed a morphosyntactic system for generating and analyzing Arabic words. The analyzer guesses the type of input word as verb, noun, or article. A little information can be used as a hint as to word type, such as a notation mark. Verb analysis starts by removing affixes and, in the case of false partitioning; it will be reattached completely or partially. After partial or full removal of affixes, the remainder will be compared with patterns that are equal in length. Entries in the pattern list are sorted according to pattern lengths and vowelization and consist of patterns, roots, and morphological properties. Many test processes are carried out to assure correct analysis including compatibility between prefix and patterns, the availability of the root in the list, and checking pronouns and syntax case of verbs.

Khoja and Garside [22] presented a linguistic algorithm that analyzes vowelized and non-vowelized Arabic words. In this algorithm, a list of function words is used to detect and filter them. The algorithm also uses backtracking as a remedy for the erroneous removal of affixes. Straightforward pattern matching is used to extract roots, where a prepared list of roots is used to check the validity of extracted roots. The algorithm tries to solve special cases, such as words containing weak letters or germination mark and hamza or glottal stop. The output of the system gives some useful statistics. The work concludes with a statement of the advantages and disadvantages of the system.

Darwish [23] designed a light Arabic stemmer that removes common prefixes and suffixes. The stemmer is very compact in size and has been significantly improved by a modification done by Leah Larkey. The stemmer accepts input in either CP-1256 or UTF-8 encoding. Internally, it transliterates input string into Roman characters, removes diacritics, and makes letter normalization. This stemmer is available publicly for research purposes.

Abuleil et al. [24] described and implemented a learning system that can analyze some types of Arabic nouns to produce their morphological information using a rule base that uses suffix analysis as well as pattern analysis. The system utilizes user feedback to classify the noun and identify the group that it belongs to. The authors classified the nouns into 84 groups according to their patterns for singular, plural, masculine, and feminine. For each group, a function was implemented to find the morphological information for a given noun. Very few groups have a unique pattern for plural and singular, and most of them share the same pattern with other groups. The system consist of the following modules: interface module, type-finder module, database module, noun morphology analyzer module, suffix-analyzer module, pattern generator module, database-checker module, and user-feedback module. It reads a noun from the text, isolates suffixes, and generates its pattern. It uses the classified noun table, the suffix/pattern analysis, or the user-feedback module to find the group to which the noun belongs and updates the database.

3.3.2 Finite-state Automata and Two-Level Morphology

Kay [25] explains that finite-state and two-level morphology has been built over the concepts of auto segmental phonology proposed by Goldsmith and auto segmental morphology proposed by McCarthy. McCarthy, as cited by Kay [25], proposed in 1979 the auto segmental theory for Arabic morphology. This work has been supported and defended by Haile and Mtenje, but Hudson raised serious theoretical challenges [26]. Beesley [27] believes that it is not

clear how McCarthy's Arabic morphology can be implemented to do practical morphological analysis or generation. Ali [28] commented that the theory developed by McCarthy might be suitable for phonology, but that applying it to morphology might be very complicated. McCarthy updated his proposal so that mosaic analysis can be applied to Arabic nouns and affixational analysis can be applied to Arabic verbs [27, 29, 30].

ALPNET and Xerox projects: In non-Semitic languages, morphemes are concatenated with roots and stems, while in Semitic language inflectional patterns are not completely concatenative but interdigitate or intercalate [31]. Stems are generated through interdigitation of roots and patterns. The standard system of Arabic orthography perhaps represents the extreme of complexity because of the great irregularities between the lexical and surface strings caused by phenomena including weak roots, hamza orthography, and the zero realization of short vowels and germination mark [27]. Even though Arabic morphology is an extreme test for any theory, a two-level model will be very suitable for the analysis of Arabic [32]. Kiraz [30] believes that nobody has dealt computationally with the challenging problem of the Arabic broken plural. Broken plurals are not handled by most of current Arabic stemmers [33].

In the ALPNET project, Beesley [27] implemented the two-level theory whereby lexicons drive the search process, while rules act as filters. In traditional two-level systems, lexicons contain morphemes, and the target is to find the correct lexicon entries that comprise the correct word. This mechanism, without modification, is not suitable for Arabic.

3.3.3 Pattern-Based Approach

Al-Kharashi and Al-Sughaiyer [34, 35] mentioned that researchers proposing different morphological analysis techniques were seeking a high degree of accuracy. This caused proposed systems to be based on heavy computational processes or the use of large amounts of associated information. Consequently, they proposed a pattern based algorithm that did not require either complex computations or associated tables. The proposed technique was based on fast-surface morphological analysis of Arabic words. Al-Kharashi and Al-Sughaiyer concluded their work as follows: Reducing the number of rules and increasing language coverage while maintaining the same level of performance and functionality can achieve enhancement. Merging rules is one method that can be used for enhancing the pattern-based stemmer. While keeping time efficiency almost the same, merging rules reduces the rule list dramatically.

4. CONCLUSION

Based on this survey we can classify the perspective of Arabic morphological systems into two major issues; one of them is the theoretical perspective and the second is the computational perspective of Arabic morphology. The computational perspective considered the Knowledge-based Approaches which consider how the algorithm deals with the Arabic word; some of these systems consider the syllable structure of the word, derivational combination of a root morpheme and a vowel melody, stem is the only morphologically relevant form of a lexeme, word forms are analyzed as arrangements of morphemes, Word-and-paradigm approach. This theory takes paradigms as a central notion, stem dependent on lexical databases, with entries associated with grammar and lexis specifications. And Empirical Approaches which deals with the rules and constraints for recognizing and classifying the internal structure of words are defined on a precise linguistic basis.

The second classification in this paper dealt with the different approaches used to build existing Arabic morphological systems. In Table lookup Approach the system depends mainly on very large tables listed in alphabetical order storing Arabic words found in natural texts with their corresponding morphological parts. While the Combinatorial Approach compares tested words against prepared lists for roots, patterns, particles, and affixes. The Linguistic Approach requires a deep analysis of the Arabic morphological system. They simulate the behavior of a linguist during the analysis of a given Arabic word. Furthermore, Traditional Approaches consider a survey of Arabic morphological analysis techniques is conducted that includes a description of the algorithm, the method of its implementation, classified by language coverage. Then the Finite-state Automata and Two-Level Morphology Approach which built over the concepts of auto segmental phonology proposed by Goldsmith and auto segmental morphology proposed by McCarthy. Finally, The Pattern-Based Approach which considers the proposed systems to be based on heavy computational processes or the use of large amounts of associated information.

References

1. Bauer, L., *The Linguistics Student's Handbook* 2007 Edinburgh University Press. 2007
2. Bohas, G., J.-P. Guillaume, and D. Kouloughli, *The Arabic Linguistic Tradition (Georgetown Classics in Arabic Languages and Linguistics series)* 2006: Georgetown University Press
3. Kiraz, G.A., *Computational Nonlinear Morphology: With Emphasis on Semitic Languages*. Studies in Natural Language Processing 2001: Cambridge University Press 194
4. Hayes, B., et al., *Linguistics: An Introduction to Linguistic Theory*. 1st Edition ed 2001: Wiley-Blackwell. 768
5. Harris, Z., *The Linguistic Structure of Hebrew*. Journal of the American Oriental Society, 1941. 62: p. 143–67.
6. McCarthy, J.A., *Formal Problems in Semitic Phonology and Morphology*, 1979, MIT.
7. McCarthy, J.A., *Prosodic Theory of Non-Concatenative Morphology*. Linguistic Inquiry, 1981. 12: p. 373–418.
8. Kiraz, G.A. *Multi-tape two-level morphology: A case study in semitic non-linear Morphology*. in *In Proceedings of COLING'94*. 1994.
9. Kiraz, G.A., *A Multi-tiered Nonlinear Morphology using Multi-tape Finite State Automata: Automata: A Case Study on Syriac and Arabic*. Computational Linguistics, 2000. 26:1: p. 77–105.
10. Kiraz, G.A., *Computational Nonlinear Morphology; With Emphasis on Semitic Languages*. Studies in Natural Language Processing, ed. I. Branimir Boguraev, T.J. Watson Research Center and L.D.C. Steven Bird, University of Pennsylvania 2004, The Edinburgh Building, Cambridge CB2 2RU, UK: The press syndicate of the University of Cambridge, the Pitt building, Trumpington street, Cambridge, United Kingdom.
11. Roark, B. and R. Sproat, *Computational Approaches to Morphology and Syntax* 2007, United States: Oxford University Press, New York.
12. Narayanan, A. and L. Hashem *Finite-state abstractions on Arabic morphology*. Artificial Intelligence Review. 1994. 7, 373–399.
13. Holes, C., *Modern Arabic: Structures, Functions, and Varieties*. Revised Edition ed 2004: Georgetown University Press. 419
14. Versteegh, K. and M.G. Carter, *Studies in the History of Arabic Grammar II: Proceedings of the 2nd Symposium on the History of Arabic Grammar, Nijmegen, 27 April-1 May 1987* 1990: John Benjamins Publishing Company. 319.

15. Cardie, C. and R.J. Mooney, *Guest editors' introduction: Machine learning and natural language*. Machine Learning, 1999. **11:1**: p. 1–5.
16. Magerman, D.M. *Statistical decision-tree models for parsing*. in *ACL-95*. 1995. ACL: Ann Arbor, MI, USA.
17. Frakes, W.B., *Stemming algorithms*. In *Information retrieval: Data structures and algorithms*1992: Prentice Hall.
18. Al-Fedaghi, S. and F. Al-Anzi. *A new algorithm to generate Arabic root pattern forms*. in *In Proceedings of the 11th national computer conference*. 1989. King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia.
19. Khawaja, A., A. Mazrui, and A.R. Boodlal, *AL-Khalil Arabic Morphological System*, 2010, Mohammed Al-Aoual University - Jeddah - Laboratory Research in informatics: The Arab League Educational, Cultural and Scientific Organisation (ALECSO).
20. Al-Bawab, M. and M.H. Al-Tayyan, *Computerized processing of Arabic morphology [in Arabic]*. Arabian Magazine of Sciences, Arab League Educational, Cultural and Scientific Organization, ALECSO, 1998. **32**: p. 6–13.
21. Al-Bawab, M., et al., *A computerized morpho-syntactic system of Arabic. The Arabian Journal of Science and Engineering*. Published by KFUPM, Dhahran, Saudi Arabia., 1994. **19**: p. 461–480.
22. Khoja, S. and R. Garside. *Stemming Arabic Text*. 1999 4-6-2012]; Available from: <http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps>.
23. Darwish, K. *Building a Shallow Morphological Analyzer in One Day*. in *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*. 2002. Philadelphia, PA, USA.
24. Abuleil, S., S. Abuleil, and M. Evens. *Acquisition system for Arabic noun morphology*. in *Association for Computational Linguistics 40th Anniversary Meeting, In Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*. 2002. University of Pennsylvania, Philadelphia.
25. Kay, M. *Nonconcatenative finite-state morphology*. in *In Proceedings of the 3rd Conference of the European Chapter of the Association for Computational Linguistics*. 1987. Copenhagen, Denmark.
26. Beesley, K.R., *Arabic stem morphotactics via finite-state intersection*, in *Paper presented at the 12th Symposium on Arabic Linguistics, Arabic Linguistic Society, 6-7 March 1998*1998b: Champaign, IL.
27. Beesley, K.R., *Computer analysis of Arabic morphology: A twolevel approach with detours* In B. Comrie & M. Eid (Eds.). *Perspectives on*

- Arabic Linguistics III, Papers from the 3rd Annual Symposium on Arabic Linguistics, Benjamins, Amsterdam, 1991. **Volume 12:** p. 155–172.
28. Ali, N., *Arabic language and computer [In Arabic]*. Ta'reeb1988.
 29. Beesley, K.R. *Arabic Morphological Analysis on the Internet*. in *6th International Conference and Exhibition on Multi-lingual Computing*. 1998a. Cambridge.
 30. Kiraz, G.A., *Computational analysis of Arabic morphology*, Computer Laboratory, George.kiraz@cl.cam.ac.uk, Editor 1995: University of Cambridge (St. John's College).
 31. Narayanan, A. and L. Hashem. *On abstract, finite-state morphology*. in *In Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics* 1993. Netherlands.
 32. Beesley, K.R. *Finite-state descriptions of Arabic morphology*. in *In Proceedings of the 2nd Cambridge Conference on Bilingual Computing in Arabic and English*. 1990. Cambridge University, UK: Literary and Linguistic Computing Centre.
 33. Xu, J., A. Fraser, and R. Weischedel. *Empirical studies in strategies for Arabic retrieval*. in *In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2002. Tampere, Finland: SIGIR 2002.
 34. Sughaiyer, I.A.A. and I.A.A. Kharashi, *An efficient Arabic morphological analysis technique for information retrieval systems*, in *In ACIDCA'2000 International Conference2000*: Monastir, Tunisia.
 35. sughaiyer, I.A.A. and I.A.A. Kharashi, *Firing policies for an Arabic rule-based stemmer*. In *Proceedings of the 9th International Symposium on String Processing and Information Retrieval (SPIRE 2002)*2002, Lisbon, Portugal.

أنظمة المحللات الصرفية والمولدات العربية الحاسوبية: دراسة استقصائية نظرية

اللغة العربية هي واحدة من اللغات التي تتميز بأن أنظمة الاشتقاق والتصريف بها من جذر واحد نظم التوابع والإعراب يمكنه إنتاج عدد كبير من الكلمات (أشكال معجمية) لكل منها نماذجها ومعانيها الخاصة تعرف بأنظمة التصريف. يعتبر علم التصريف من وجهة نظر اللغويين هو دراسة بنية الكلمات. وبعبارة أخرى، التصريف هو تعبير عن ذلك الفرع من علم اللغة المعنية بالأشكال والأبنية التي تأخذها الكلمات لاستخدامات لغوية مختلفة. لذا فيعتبر نظام تحليل التصريف أحد أهم الأجزاء في كافة نظم معالجة اللغات الطبيعية كرامج الترجمة الآلية وبرامج استرجاع المعلومات.

في حين يتم بناء العديد من أنظمة الصرفية العربية مع استراتيجيات ومنهجيات مختلفة؛ تعتبر هذه الورقة دراسة نظرية للأنظمة الصرفية العربية من حيث وجهات نظرهم الطرق المختلفة المستخدمة في بناء الأشكال الصرفية للكلمة العربية. فبناء على هذه الدراسة يمكن أن نصنف وجهة نظر النظم الصرفية للغة العربية في قضيتين رئيسيتين، إحداهما هي الناحية النظرية، والثانية هي وجهة النظر الحاسوبية. فمن الناحية النظرية يمكن اعتبار الأجزاء الصرفية للكلمة هي التي تتضمن دلالات لغوية يمكن كتابتها بأحرف اللغة. ومن وجهة النظر الحاسوبية فمنها التي تعتبر الطرق القائمة على قواعد البيانات والتي تعتبر الكيفية التي تتعامل بها مع خوارزمية الكلمة العربية، فبعضها إما يعتمد على نظرية هيكل المقاطع الصوتية من الكلمة، أو نظرية التوافقات الاشتقاقية من جذر الكلمة أو أصلها. وهناك من النظريات التي تعتمد أن الجذر هو الشكل الوحيد شكليا ذات الصلة بالكلمة، والبعض الآخر يعتمد تحليل أشكال الكلمة من كافة أشكالها المعجمية أو التي تعتمد على قواعد البيانات المعجمية، أو مع الإدخالات المرتبطة بمواصفات النحو. أو البنية الداخلية للكلمات على أساس لغوي دقيق.

أما الجزء الثاني من هذه الورقة فيعرض الطرق المستخدمة لبناء نظم الصرف العربية نفسها. فمنها طرق الجداول البحثية Table Lookup Approach، أو الطريقة التوافقية Combinatorial Approach، أو اللغوية Linguistic Approach، أو الأساليب التقليدية Traditional Approaches، وغيرها. ففي طرق الجداول البحثية يعتمد النظام بشكل رئيسي على جداول كبيرة جدا تعتمد على الترتيب الأبجدي لتخزين الكلمات العربية الموجودة في النصوص الطبيعية مع أجزائها الصرفية المناظرة. في حين أن الطرق التوافقية تعتمد على مقارنة الكلمات بقوائم معدة سلفاً لهذه الجذور، وأوزانها، وأشكالها، ولاحقاتها. أما الطرق اللغوية فتطلب تحليلاً دقيقاً للنظام الصرفي العربي، باعتبارها محاكاة للسلوك اللغوي في تحليل كلمة بعينها. وعلاوة على ذلك، فإن النظر في الطرق التقليدية يجري مسحا لتقنيات التحليل الصرفي العربي الذي يتضمن وصفا للخوارزمية، وطريقة تنفيذه، مصنفة حسب التغطية اللغوية. وهناك من الطريق ما يسمى بطريق التشغيل الذاتي المحدودة Finite-state Automata and Two-Level، وذات المستويين بنيت على مدى مفاهيم علم الأصوات. وأخيراً، فإن هناك الطرق القائمة على النمط أو الوزن والتي تعتبر النظم المقترحة قائمة على أساس العمليات الحسابية أو استخدام كم هائل من المعلومات المرتبطة بها.

1870

1871

1872

1873

1874

1875

1876

1877

1878

1879

1880

1881

1882

1883

1884

1885

1886

1887

1888

1889

1890

1891

1892

1893

1894

1895

1896

1897

1898

1899

1900

1901

1902

1903

1904

1905

1906

1907

1908

1909

1910

1911

1912

1913

1914

1915

1916

1917

1918

1919

1920

1921

1922

1923

1924

1925

1926

1927

1928

1929

1930

1931

1932

1933

1934

1935

1936

1937

1938

1939

1940

1941

1942

1943

1944

1945

1946

1947

1948

1949

1950

1951

1952

1953

1954

1955

1956

1957

1958

1959

1960

1961

1962

1963

1964

1965

1966

1967

1968

1969

1970

1971

1972

1973

1974

1975

1976

1977

1978

1979

1980

1981

1982

1983

1984

1985

1986

1987

1988

1989

1990

1991

1992

1993

1994

1995

1996

1997

1998

1999

2000

2001

2002

2003

2004

2005

2006

2007

2008

2009

2010

2011

2012

2013

2014

2015

2016

2017

2018

2019

2020

2021

2022

2023

2024

2025

2026

2027

2028

2029

2030

2031

2032

2033

2034

2035

2036

2037

2038

2039

2040

2041

2042

2043

2044

2045

2046

2047

2048

2049

2050

2051

2052

2053

2054

2055

2056

2057

2058

2059

2060

2061

2062

2063

2064

2065

2066

2067

2068

2069

2070

2071

2072

2073

2074

2075

2076

2077

2078

2079

2080

2081

2082

2083

2084

2085

2086

2087

2088

2089

2090

2091

2092

2093

2094

2095

2096

2097

2098

2099

2100

Enhancing UML Conceptual Modeling Through The Use Of Integration

*Farouk N.**

*Salah A.**

*Kamel A.**

Abstract

It is currently recognized that the Unified Modeling Language (UML) tends to be the most powerful and commonly used software modeling and visualizing tool in the software engineering industry. UML tools are used to simplify the complexity of the software systems and facilitate analyzing and designing them, they do this by breaking the system into small parts or views. Unfortunately, considering the system as a set of views makes UML diagrams have no explicitly relationships between each other. The developer always makes these relationships on his mind to generate one big picture for the software system. This research proposes a framework for enhancing the modeling process quality by generating an integrated view for UML diagrams. This framework works on use-case, sequence, state chart and class diagrams. Integration is a powerful way for understanding the whole view of the software system by defining the semantic relationships among UML diagrams. The proposed integrated model provides a means for avoiding conflicts in the software system that appear due to mismatches between two views within UML. This is particularly important during making changes in the software system and even more important as the software system size or the team size increases.

Keywords: Software Visualization, Conceptual Model, Integrated Model, Unified Modeling Language UML, Software Development Process.

1. Introduction

Understandability of software is one of the most important factors of the quality of the software system. Understandability can be increased by maintaining extra documentation. But, in most cases the only regularly updated documentation is the source code itself and the understandability depends a lot on its quality [8]. Software Modeling becomes the most activity that has been carried out over the years in software development process. When developing system by using the simplest languages to the most powerful and complex languages, the developer still needs to model. Defining a model makes it easier to break up a complex application or a huge system into simple, discrete parts that can be individually studied. It can be focused more simply on the smaller parts of a system and then understand the whole picture [2].

The Unified Modeling Language UML now becomes one of the major conceptual modeling tools in the field of Object Oriented software engineering [3], [4]. UML is defined to be the language that can be used to model software systems and make them readable and understandable. Thus UML provides the ability to capture the characteristics of a software system by using special notations. UML provides a wide range of simple, easy to understand notations for documenting systems based on the

* Computer Science Department, Faculty of Computers and Information, Cairo University, Egypt

Object-oriented design principles. UML has the advantages of easier designing, larger reusability, and higher quality [9].

UML diagrams that describe any software system can be divided into two distinct characteristics: a structural, '*static*' part which is mainly the structural aspect of the system. The static characteristics define what parts the system is composed of. And a behavioral, '*dynamic*' part which describes the behavioral features of a system; the ways a system behaves in response to certain events or actions. The diagrams that describe the Static part are; the Use-case Diagram and, the Class Diagram. Also the diagrams that describe the Dynamic part are; the Object diagram, the State diagram, the Activity diagram, the Sequence diagram and, the Collaboration diagram [2].

As UML has advantages, it has some drawbacks that are explored recently. As mentioned before, the UML is used to simplify the complexity of the software systems, but dealing with this complexity has been done by breaking the system up into discrete parts and considering the system from particular perspective or view. Views enable software developers to reduce the amount of information they have to deal with at any given time. *"It has been recognized that it is not the number of details, as such, that contributes to complexity, but the number of details of which we have to be aware at the same time"* [8]. Each software developer, involved in the software team, needs to view the software system from his point of interest only or the point that is important to know, to do the role that is based on him in the development process. Considering software systems as collection of views makes the software development process tends to become 'view centric' which means, seeing particular views instead of the big picture. This view centric approach causes a reasonable amount of redundancy among the different views as a side effect. That redundancy is the cause for inter-view mismatches, such as inconsistencies or incompleteness. On top of that, while the software size or the team size increases, it becomes harder to manage the software development process. As everyone involved in the team needs to know the changes that happened for the design of the software system because this change might make him change in the part he is responsible for. Views are used independently, concurrently, are subjected to different audiences and the manner in which model information is shared is extremely inconsistent. All this implies that information about a system must be captured multiple times and must be kept consistent [1].

This research proposes framework architecture for integrating some UML diagrams, as a start, to enhance the software development process, and help the software developers to facilitate dealing with the model of the software system. Although they have to consider the software system model as a collection of views, they also can consider the whole picture of the software system model.

The rest of the paper is structured as follows; Section (2) introduces the used UML diagrams and the relationships between them. Section (3) presents the proposed framework architecture for the integration process. Section (4) discusses the consequences benefits of the integration process. In section (5), conclusion and future work.

2. UML diagrams

As mentioned before that UML diagrams can be divided into two characteristics (static and dynamic). The research here takes in consideration these types of characteristics. The diagrams that the proposed implemented tool started with are; 'Use-case diagram', 'Class diagram' which represent structural or static part, and 'Sequence diagram', 'State chart' which represent the behavioral or dynamic part. Let's take a deep look into these diagrams and what do they represent [5];

- **Use-case diagram:** the use-case diagram is used to identify the primary elements and processes that form the system. The primary elements are termed as "actors", might be a user or an external system, and the processes are called "use cases". The use case diagram shows which actors interact with each use-cases.

- **Class diagram:** The class diagram is used to refine the use case diagram and define a detailed design of the system. The class diagram classifies the functionalities defined in the use case diagram into a set of interrelated classes. The relationship or association between the classes can be either an "is-a" or "has-a" relationship. Each class in the class diagram may be capable of providing certain functionalities. These functionalities provided by the class are termed "methods" of the class. Apart from this, each class may have certain "attributes" that uniquely identify the class.

- **Sequence diagram:** A sequence diagram represents the interaction between different objects in the system. The important aspect of a sequence diagram is that it is time-ordered. This means that the exact sequence of the interactions between the objects is represented step by step. Different objects in the sequence diagram interact with each other by passing "messages".

- **State chart:** A state chart diagram, as the name suggests, represents the different states that objects in the system undergo during their life cycle. Objects in the system change states in response to events. In addition to this, a state diagram also captures the transition of the object's state from an initial state to a final state in response to events affecting the system.

2.1 Specifying the semantic relationships and dependencies between UML diagrams

Different UML diagrams represent different system aspects, but their coupling is not precisely defined, i.e. even if one had a formal semantics of every employed diagram type it would not be always clear in which way the diagrams and their semantics interact [7] [1]. This section describes the semantic relationships between the four used UML diagrams and each others. This helps understanding the correlations and the interactions between these diagrams.

2.1.1 Relationship between Use case diagram and Sequence diagram

Use case diagrams shows the goals that various actors have in interacting with the system. But a use case diagram doesn't give any details about how those goals are achieved. Because sequence diagrams represent participant interactions over time the developer can use sequence diagrams to show how a use case is executed. One sequence diagram typically represents the flow of events for one individual use case. Sequence diagrams are used to document use case scenarios. Sequence diagrams may help the developer finding necessary interactions that the written use case didn't capture [6].

2.1.2 Relationship between Class diagram and Sequence diagram

When the class diagram used in conjunction with sequence diagrams, it provides an extremely effective communication mechanism. The developer can use a class diagram

to illustrate the relationships between the classes, and the sequence diagram lets the developer show the messages sent among the instances of these classes and the order in which they are sent. When an object sends a message to another object, it implies that the two classes have a relationship that must be shown on a class diagram [6].

2.1.3 Relationship between Use case diagram and Class diagram

As many modelers consider class diagrams are the backbone of their overall model of a system. A class diagram describes the kinds of objects that populate the system as well as the kinds of static relationships that exist among those classes and objects. Class diagrams are often used to refine use case diagrams. Use case diagrams model the system's functionality from a perspective outside the system. As such these diagrams identify the primary elements and processes that make up the system [6].

2.1.4 Relationship between Sequence diagram and State chart diagram

The relationship between sequence diagrams and state chart diagrams can bind states in state chart diagrams with messages in sequence diagrams. In the sequence diagrams, objects interact with other objects by exchanging messages, it shows a composition of objects exchanging messages, without describing the internal behavior of these objects. The state chart diagrams show a set of states that an object can take. The transition from one state to another state is triggered by an event, which can be a received message. And the description of this message is showed in sequence diagrams [6].

Now, after defining some UML diagrams semantic relationships, it has been clearly found that the developer should recognize the software system from all perspectives. That, showing the related diagrams with each other helps the software developer sees the diagram element from more than one side. This also helps the developer while changing any element or diagram to recognize the related elements or diagrams that are related to be changed too, if required.

3. Proposed Framework

As we mentioned before, the work here will take in consideration a four UML diagrams (Use-case diagram, Class diagram, Sequence diagram and State chart). The purpose of the work here is to find a way to integrate these UML diagrams and find the relationship between these diagrams and each other. Let's take an example; assume a class called "X". This class is involved in many diagrams of the software UML model like "Class Diagram1", "Sequence Diagram1", "Sequence Diagram4", "Sequence Diagram6", and has a State chart called "State chart 5". It is important to the developer to know the states of that class, its contents, structure and the relationship between this class and the remaining classes. The developer may need to open each diagram and then search if that class exists or not in the opened diagrams, there is no other choice, and then keeps these related diagrams into his mind every time he wants any information about this class.

By presenting the "Integrated UML" tool here that provide the integrated UML model, when the developer press on the Class "X" in the Class diagram, a list appears that contains other related diagrams ("Sequence Diagram1", "Sequence Diagram4", "Sequence Diagram6", "State chart1") that class "X" is involved. This will provide viewing the whole picture of class "X"; its structure, contents, and relationship between

that class and others and the data-flow of that class. Also the developer can navigate through these diagrams and open any of them.

3.1 The Framework Architecture

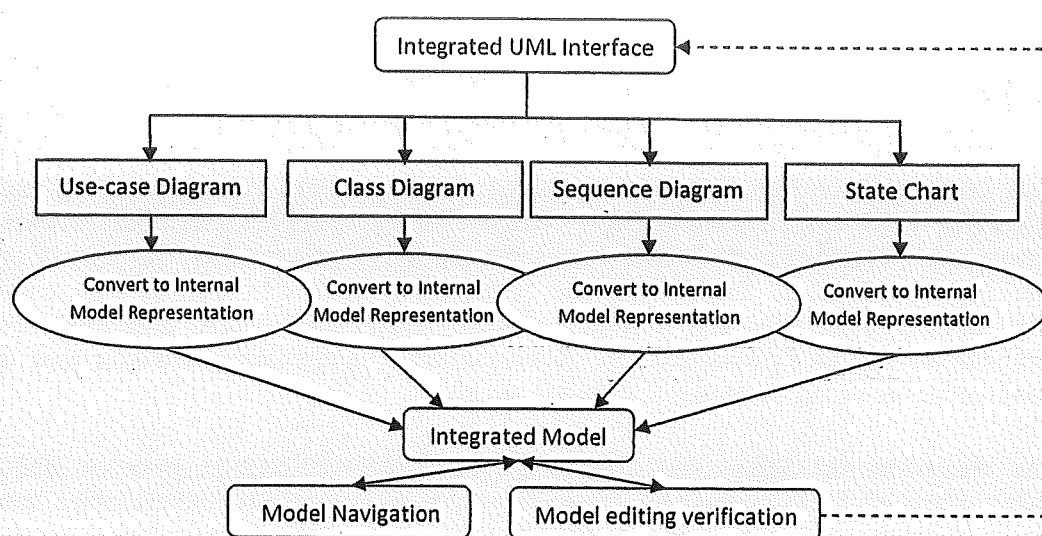
The framework presented in this research has the following components; (Figure-1)

- **Integrated UML Interface**

The Integrated UML program Interface contains the used UML diagrams; Use-case Diagram, Sequence Diagram, Class Diagram and State-chart Diagram. These diagrams will be input to the implemented tool file by file for each diagram to be viewed in the screen.

- **Diagram Conversion into Internal Model Representation**

The diagrams files and the objects that are involved in each diagram will be stored into a relational database structure to facilitate dealing with (read, write and search).



Figur-1. Proposed Framework

- **Integrated Model**

The objective of the work itself is to provide an integrated model contains the UML diagrams and their relationships to present the overall view of the software project model. The developer will put the relationships between diagrams and each others, and the relationships between objects in those diagrams into the database once. Then among the development process, the developer can show those relationships and find which diagrams, objects related to. The integrated model will facilitate interaction between the software project team and each other that each one of them will see the whole view of the software model and can be aware of any changes made.

- **Model Navigation**

The integrated model provides navigating form one diagram to another in simple way, that makes the software project team aware of the relations between the UML diagrams and each other, and also the objects inside the diagrams and each others.

- **Model Editing and Verification**

The integrated model also provides verifying the model diagrams and the consistency of these diagrams. Model verification will help the developer while doing any changes in one diagram with a list of related diagrams to change too (if required).

3.2 Assumptions and Constraints

- The “used UML diagrams” are the UML diagrams that the implemented tool will start working on, and they are; Use-case diagram, Sequence diagram, Class diagram, and State chart. The used UML diagrams are the most commonly used diagrams that are made by software developers in the process of developing software program. In future, more diagrams can be added that will be more useful in the process of developing software programs.

- In the process of converting the diagrams into some data structure, we assume that the input diagrams are correct and verified.

- The data structure that is used in the implemented tool is a relational database. The use of database concept could facilitate storing, updating and retrieving the details of diagrams and the objects' relationships.

3.3 Integration Methodology

The Obvious way to integrate the used UML diagrams is to transform these diagrams from graph-level into another lower level of abstraction that provides dealing and interacting with them. Thus in the graph level, it is hard to make the integration between the diagrams because in this visual representation by graphs we have no criteria of checking dependencies, consistencies and correctness. The lower level of abstraction will facilitate recognizing and storing the inter-relationships between diagrams. The backbone of the tool presented here is the Internal Model Representation that is provided in some kind of relational database structure.

As we mentioned before that there are a semantic relationships between UML diagrams and each others, now let's define a way to represent these relationships into relational database; as the developer should define at first the relationships between diagrams, objects and each others for the software system only once. Some of these relationships might be recognized as follow;

- For each use-case in the use-case diagram there is a sequence diagram,
- Each class in the sequence diagram should exist in the class diagram,
- Each class in class diagram may have a state chart,
- Each class in the state chart should exist in the class diagram and in some sequence diagrams.

Providing the integrated model will be simple by taking in consideration the above relationships (assuming that the database already filled by data). Thus while clicking on any object at any diagram in the implemented tool interface; the program automatically makes a search query to find the list of the related diagrams that contain the same object, and the developer can choose to easily transfer to one of these related diagrams. Thus the developer can see the Class in the class diagram, know the properties and functionalities, see the relationships with the other classes, and also find that Class in the related Sequence diagrams and see the messages pass from and to it from other classes, and find what use cases that the sequence diagrams represent, know the goal of this use case and which actors deal with it, and also find the State chart of that class and recognize the states of that class that are changed through events.

By the above scenario, the user can navigate through the model and have a complete overview about the software system model, integrated model, by exploring the related diagrams that represents more than one perspective of the software system. The consistency between the diagrams is provided by using the internal representation model that can check what are the related diagrams that the user has to change them while changing in any diagram.

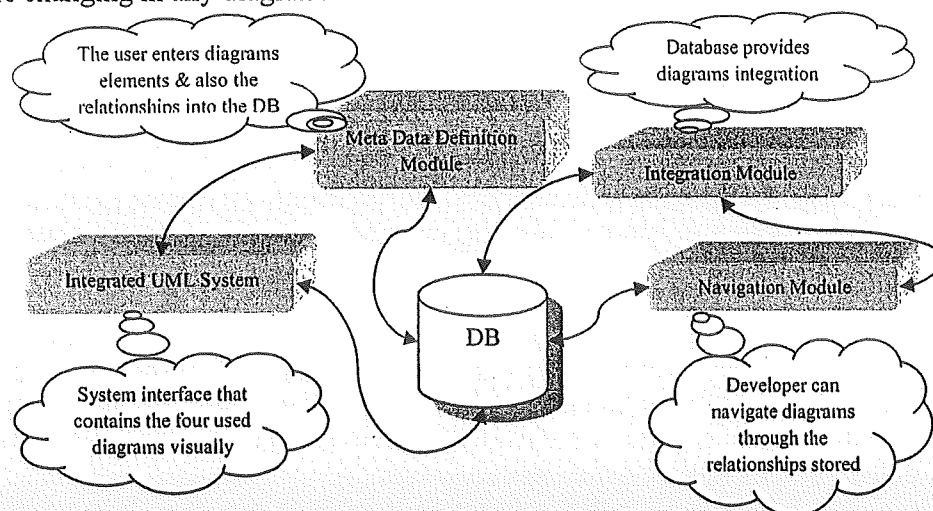


Figure-2. Integrated UML Modules

Figure-2 illustrates the "Integrated UML" tool modules. A prototype was built to test this architecture using C# for the system interface; it is a windows application, Microsoft Visio for viewing the UML diagrams in visual manner and Microsoft Access database to store the diagrams into data structure manner.

4. Benefits of Providing an Integrated Model

The advantages of representing multiple views of UML diagrams in the process of integration are; a common data model that represents the overall view of the software system and, a common set of tools that can be used to manipulate that integrated model. Integration extends representation by also describing interactions among multiple views, thus capturing the inter-view relationships. Those inter-view relationships are essential to enable automated identification of consistency and conformance mismatches [1].

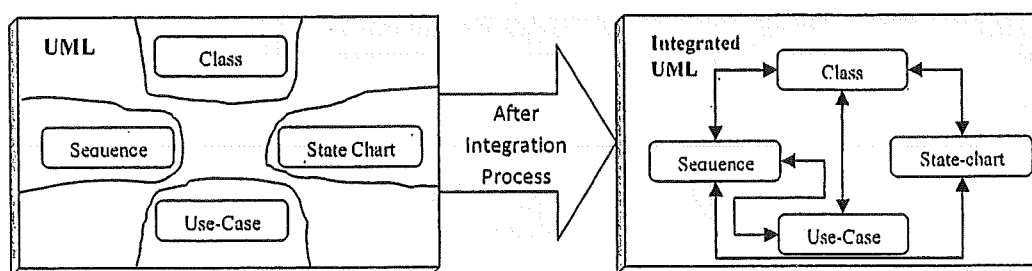


Figure-3. UML diagrams before and after Integration process

Figure-3, the left side box represents the used UML diagrams that are independent, not related to each other. Each one of them represents a separate view of the software system. The right side box represents the used UML diagrams after integration process

and provides the overall view of the software system. Also represents the inter-relationships between the diagrams and each other. The advantage of considering the inter-relationships between diagrams is that these diagrams are connected by mutual dependencies and relationships that have to be kept consistent with each other. Thus, if one diagram is changed, these changes have to be propagated to dependent diagrams in order to restore mutual consistency. The integration will provide the consistency, verification and correctness of the diagram.

5. Conclusion and Future Work

The work here illustrated the differences between view representation and view integration. The research introduced framework architecture for integrating use-case, sequence, state-chart and class diagrams by defining the dependencies and the relationships among these diagrams and each other. Integration process enhances the quality of modeling and increases the value of the UML model diagrams in the software development process by presenting the objects, diagrams with all perspectives for the developers. The proposed framework illustrates the advantages of using the integrated UML diagrams over the process of software development while the system size or the number of developers in the software team increases.

The next step is to complete the implementation of the proposed framework and include the remaining UML diagrams. Many things remain to be done, both at conceptual level or implementation level. The proposed framework could be applied on a multi-user system that the team developers can be connected to each other and see the UML diagrams at the same time and show the changes applied from each other.

References

- [1] A. E. a. N. Medvidovic, "Extending Architectural Representation in UML with View Integration," in *the 2nd International Conference on the Unified, Fort Collins, October 1999*.
- [2] P. T. & L. A. Mandar Chitnis, "<http://www.developer.com>," 9 December 2002. [Online].
- [3] K. L. Siau, "Theoretical and Practical Complexity of Unified Modeling Language: Delphi Study and Metrics Analyses," in *International Conference on Information Systems*, 2004.
- [4] B. W. Simon M. Becker, "Uml-Based Definition Of Integration Models For Incremental Development Processes In Chemical Engineering," 2003.
- [5] Yves Dumond, Didier Girardet, Flavio Oquendo , "A Relationship Between Sequence And Statechart Diagrams".
- [6] N. Conner, "Virtual Training Company- VTC," 26 October 2007. [Online]. Available: <http://www.vtc.com/products/UML/TypesofDiagramsAnOverview/66827>.
- [7] C. Ermel, "Animated Simulation of Integrated UML Behavioral Models based on Graph Transformation".
- [8] H. R. F. S. Hans-Gerd Köhler, "Understanding Object Oriented Software Systems Without Source Code Inspection," Software System Engineering Group, Technical University of Cottbus, Germany.
- [9] Bin Zhang and Ye-sho Chen, "Enhancing Uml Conceptual Modeling Through The Use Of Virtual Reality," in *Proceedings of the 38th Hawaii International Conference on System Sciences*, 2005.

Human like Query Tool: Non-Stationary Fuzzy Set Approach

Sabour. A. A.** , Rahman. B. M.** , Tawab. A. H.** , Gadallah. A. M.* , and Hefny. H. A.*

Abstract

Users often have vague or imprecise ideas when searching databases and thus may not know how to precisely formulate queries that lead to satisfactory answers. Also, the user searching criteria may depend on the current state of the data stored in the searched database. Based on fuzzy set theory and non-stationary fuzzy sets, this paper proposes a fuzzy query approach supporting database – sensitive fuzzy queries in which the definition of a non-stationary fuzzy set is dynamic and sensitive to the current database state. The proposed approach enables the user to generate query statement that may include non-stationary fuzzy operands and operators. Accordingly answers of such queries become more sensitive to the current database state and more satisfactory.

Keywords : Fuzzy Set, non-stationary Fuzzy Set, SQL, Fuzzy Query, Linguistic variable, Linguistic value, Fuzzy Logic, Fuzzy Relational Database.

1 Introduction:

FUZZY sets were first introduced by Zadeh [1] to model the uncertainty inherent in assigning memberships of elements to real-world sets such as the set of old people or the set of tall people. They were specifically designed to represent vagueness, and provide formalized technique for dealing with imprecision in real-world problems. Generally speaking, fuzzy sets are used to enhance classical queries to make it more closed to human-like queries. In practice, however, many

*Computer Science Dept., Institute of Statistics, Cairo Univ.

**MSc Student, Information Systems Dept., Institute of Statistics, Cairo Univ.

complex decision-making problems are expressed using classical fuzzy sets. In some situations, representing a linguistic value likes “excellent students” using a classical fuzzy set is not efficient when applied to different database states with a high variance among them. In other words, if the average of student grades in one database state is 80%, then this may imply that “Excellent Students” are students with grade around 95%. On the other hand, if the average grades 73%, then this may imply that “Excellent Students” are student with grade around 87%.

In general, fuzzy queries mean that the query condition contains fuzzy terms including fuzzy values, fuzzy hedges, fuzzy numbers and fuzzy operators. This work enables the user to define his own fuzzy terms reflecting his/her point of view. Also, such fuzzy terms can be defined using non-stationary fuzzy sets indicating its sensitivity to the current state of the database. In other words, the control points of the used membership functions may be derived from the current database state. Accordingly, a user should be able to define his/her own fuzzy terms respecting the nature of each fuzzy term. That is for defining some fuzzy terms he/she can use traditional stationary membership functions and for defining other fuzzy terms he/she can use non-stationary fuzzy sets. Consequently, when a user generates a fuzzy query using non-stationary fuzzy sets, then both of query and result will be sensitive to the current database state. Finally, the result will include only tuples with matching degree greater than or equal the defined threshold. Also, the resulted tuples will be ranked according to their matching degree to the original fuzzy query.

The rest of this paper is organized as follows: section 2 overviews basic concepts of fuzzy logic. Related works were introduced in section 3. Section 4 presents the proposed fuzzy query approach architecture. The proposed tool and an illustrative case study were given in sections 5. Section 6 concludes the paper.

2 Fuzzy Concepts Overview

This section introduces an overview of basic concepts of Fuzzy Logic

2.1 Fuzzy Logic

Fuzzy set theory is the basic building block of fuzzy logic. In this logic, the truth-value of a sentence (or satisfaction degree) is in the real interval $[0, 1]$. The value 0 represents completely false, and 1 is completely true. The truth-value of a sentence "s" will be denoted as $\mu(s)$. Fuzzy logic was developed as a mean to do reasoning under uncertainty. Many new approaches and theories treating imprecision and uncertainty have been proposed since fuzzy set was introduced by Zadeh [15].

2.2 Fuzzy Set

Fuzzy sets are defined on a non-fuzzy universe of discourse, which is an ordinary set. A fuzzy (sub) set F of a universe of discourse U is characterized by a membership function $\mu_F()$ which assigns to every element $x \in U$, a membership degree $\mu_F(x) \in [0, 1]$. An element $x \in U$ is said to be in a fuzzy set F if and only if $\mu_F(x) > 0$ and to be a full member if and only if $\mu_F(x) = 1$. Fig. 1 shows membership functions of fuzzy sets "very young", "young", "old", and "very_old" defined on the universe of age.

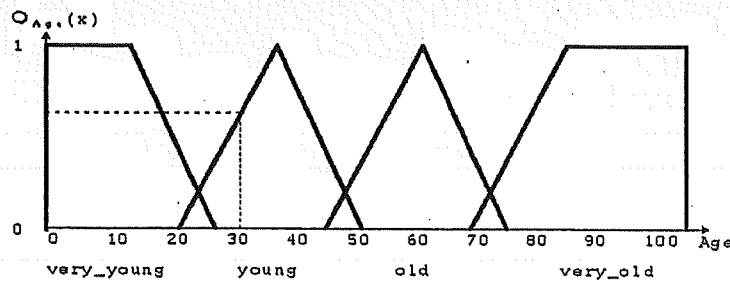


Figure 1: membership function for AGE attribute.

2.3 Fuzzy Truth Values

A fuzzy set operation is an operation applied on fuzzy sets. These operations represent a generalization form of crisp set operations. There is more than one possible generalization form for each fuzzy operation. The most widely used operations are called standard fuzzy set operations. The return value can be viewed as a number in $[0,1]$. The value 0 corresponds to false of Boolean logic while the value 1 is equivalent to true. These operations are defined as shown in table 1.

Table 1: Fuzzy Truth Values[16]

Operation	Name	Definition
Complement	not	$1 - x = \bar{x}$
Fuzzy and	and	$\min(x, y) = x \& y$
Truncation	mod	x if $x > y$, otherwise 0
Fuzzy or	or	$\max(x, y) = x \vee y$
Rounding	ren	x if $x < y$, otherwise 1
Fuzzy xor (distance)	xor	$\max(\min(x, 1 - y), \min(1 - x, y)) = (x \& \bar{y}) \vee (\bar{x} \& y)$
Distance	-	$ x - y $
Probabilistic or	-	$x + y - x \cdot y$
Probabilistic and	*	$x \cdot y$

2.4 Fuzzy Query

Fuzziness is introduced in database query language to allow imprecise querying on conventional data. If we aim for conventional crisp querying languages to allow precise and certain querying for traditional Databases. Precision assumes that the effect will be exactly our perception and certainty assumes the structure and parameter are exactly identified. But for factual database there may be some complications. Actual situations are very often not crisp and

deterministic and cannot be described precisely i.e. Real situations are very often uncertain or vague in a number of ways. Also, complete description of a real system would entail far more detailed data than a human being could ever be acquainted with and process at the same time. To get rid of these complications we need to be concerned about the notion of uncertainty. An analogous example would be our everyday SQL selects. Let's say a company is searching for young employees with great sales records, to consider for promotions.

This query statement may be constructed as follows:

```
SELECT *  
FROM   employee  
WHERE  age BETWEEN 20 AND 30  
AND    sales > 100000
```

This statement will miss the 21 years old with \$99,000 in sales and the 31 years old with \$500,000, although those people may be bright young stars as well. What the company actually wants is to do a SQL statement. Selecting young employee with great sales record. One solution would be fuzzy logic. They want employees that fall into two sets: 1) Young and 2) Good sales.

The fuzzy solution would say, ok, every employee selling over \$100,000 is 0.6. Also, anyone less than 30 years old is a member of the Young Employee set with membership 1.0. 31 years old is 0.8. 35 years old is 0.5. Once you define those parameters, by definition the membership of an employee in the two sets is the lowest membership he has in either set. Our precocious 21 year-old would be 0.8 (he has 1.0 in Young, and 0.8 in Good Sales) and our older but productive 31 year old would also be 0.8 (0.8 Young and 0.8 Good Sales).

3 Fuzzy Inference System And Non-Stationary Fuzzy Sets

Commonly, a fuzzy inference system (FIS) consists of four main interconnected components: the rules, the fuzzifier, the inference engine, and an output processor. Type-1 FISs use only type-1 fuzzy sets, whereas an FIS that uses at least one type-2 fuzzy set called a type-2 FIS. Fig. 2 shows the mechanisms of a type-2 FIS. Fig. 3 shows the mechanisms of the inference process in a FIS consisting of non-stationary fuzzy sets. We naturally term such an FIS a “non-stationary FIS”. It should be emphasized that a non-stationary FIS is simply a repetition of a type-1 FIS with slightly different instantiations of the membership functions over time. Thus, implementing a non-stationary FIS is simply a matter of iterating over the required number of instantiations while perturbing the membership functions. Neither the form of non-stationarity (variation in location, variation in width, or noise variation) nor the form of perturbation function (periodic, random, chaotic, etc.) has any effect on the inference process. Hence, an inference with non-stationary fuzzy sets is clearly different from the type-2 inference, and does not suffer the difficulties of type-2 inference (particularly the inference using general type-2 fuzzy sets).

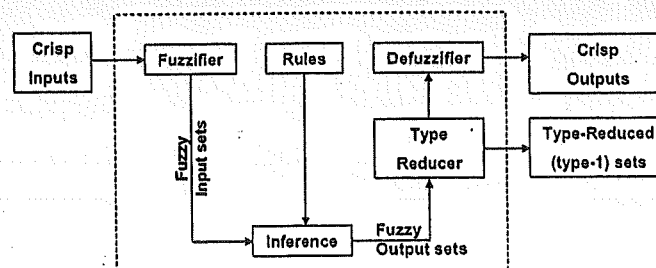


Fig. 2. Mechanisms of a type-2 FIS

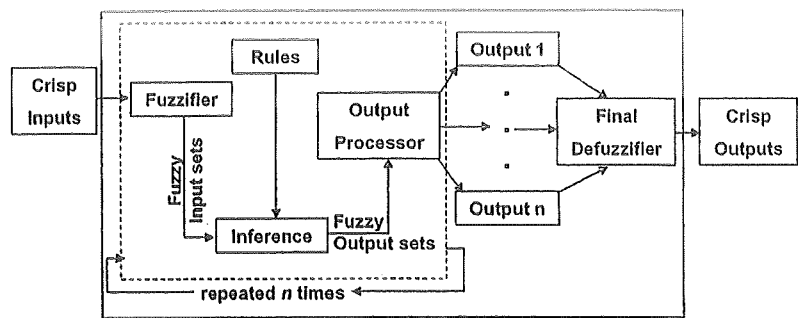


Fig. 3. Architecture of non-stationary FIS.

4 The proposed architecture

In this section, we present the new proposed architecture to model the flexible fuzzy queries.

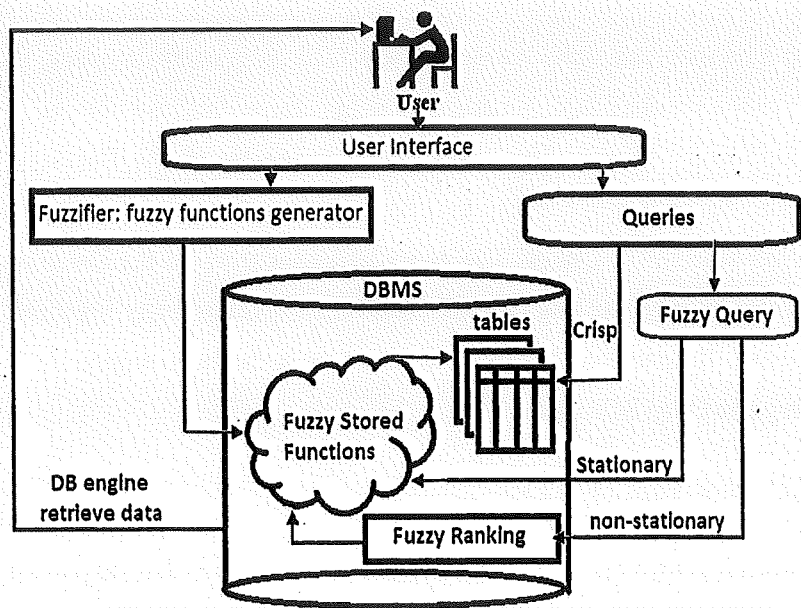


Fig. 4: The architecture of the proposed Fuzzy query

The processing scenario of the proposed fuzzy query approach

Working principle of all components of the above proposed architecture is given below:

- **User Interface:** users can easily define his/her own fuzzy profile in a flexible way manner that includes the definition of user's fuzzy terms. Preferences are user oriented that works according to the settings set by each user separately (User Profiling).
- **Fuzzifier:** The program begins by creating PL-SQL Statements that creates database stored objects (such as packages, procedures, functions, views, etc.) within the database, these objects may encapsulate the user defined data needed to define their behaviors instead of using Fuzzy Meta Knowledge Base (FMB), in order to accelerate execution and lifting performance and efficiency when performing fuzzy queries.
- **Queries:** The user can write queries or use GUI to generate a query, the query is identical to the standard SQL language by using stored database objects.
 - **Stationary Query:** The query will go directly to the Database engine to execute directly in one step as it does with regular Crisp SQL statements.
 - **Non-stationary Query:** The query will go directly to the Database engine to make "Fuzzy Ranking" to tuples and then execute fuzzy functions to get the result. Responsible for all of these steps is database engine as it does with regular Crisp SQL statements.

5 Fuzzy Query Tool (FQT)

In this section, we present the querying language FQT that apply the proposed approach. The proposed FQT includes a set of modules

5.1 Linguistic Variables

This module allows create, modify, or delete linguistic variables owed by the user. Just give a name for a linguistic variable then click on the Create or Replace button as shown in Fig. 5.

FQPT Fuzzy Query Processing Tool  

Welcome Back: Fuzzy [main](#) [logout](#)

Linguistic Variables Page

Linguistic Variable:

Remove	OWNER	linguistic variables	STATUS	TIMESTAMP
Remove	FUZZY	AgeDegree	VALID	2012-06-11:17:15:10
Remove	FUZZY	AGEDEGREE	VALID	2012-06-11:17:19:08
Remove	FUZZY	AgeM	VALID	2012-06-11:19:30:34
Remove	FUZZY	AGEM	VALID	2012-06-11:19:32:34

Fig. 5: Linguistic Variables Screen

5.2 Non-Stationary Fuzzy Term Module

A non-stationary fuzzy set has a dynamic definition instead of a fixed one like traditional fuzzy sets. It may be sensitive to the state of the environment as shown in Fig. 6 and Fig.7.

FQPT Fuzzy Query Processing Tool  

Welcome Back: HR [main](#) [logout](#)

Non-Stationary Linguistic Values

Linguistic Variable:

Linguistic Value:

Function Type:

☒ Triangular
 ☐ Trapezoid
 ☐ Gaussian
 ☐ Gaussian bell-shaped

Select	Variable	Values	Function	Percentage	Operator	Value	Order
☒	SAL	HiSal	Triangular	35	>	9000	1
☒	SAL	LowSal	Triangular	70			2

Linguistic Values Saved

Language: en-us Copyright © 2011, 2012, ABA - GadAllah. All rights reserved.

Fig.6: Defining a non-stationary linguistic values

Non-Stationary Linguistic Values

Linguistic Variable:
 Percentage
 Value as a percentage:
☒ with condition
 Conditions and boundaries has first priority
 Condition Operator: Value:
 Linguistic Value:
 Function Type:
☒ Triangular ☐ Trapezoid ☐ Gaussian ☐ Gaussian shaped ☐ bell-

Select	Variable	Values	Function	Percentage	Operator	Value	Order
>>	GRADES	Excellent	Triangular	10			1
>>	GRADES	VeryGood	Triangular	15			2
>>	GRADES	Good	Triangular	30	>	70	3
>>	GRADES	Pass	Triangular	35	>=	50	4
>>	GRADES	Failed	Triangular	40	<	50	5

Fig.7: A set of non-stationary linguistic values defined over a linguistic variable

Non-Stationary fuzzy set indicates a fuzzy set with dynamic control points instead of fixed values. Such dynamic control points are determined based on some concluded values from the current database state. The user can also define some linguistic values to be 'High' and 'Low'. The fuzzy set of 'High salaried employees' includes the first 35% highest salaried employees. On the other hand, the rest of employees represent the fuzzy set 'low salaried employees'. Accordingly such fuzzy sets are defined to be non-stationary because of their dependency on the current database state. The following example presents how to use non-stationary linguistic values defined as non-stationary fuzzy sets in constructing fuzzy query statements, see Fig. 7 and Fig.8.

```
select
salary,Rnk,Percentage,SallngVal(Salary,Percentage)
from employees_salary_view
```

Run Query

SALARY	RNK	PERCENTAGE	SALLNGVAL(SALARY,PERCENTAGE)
24000.00	107	1	HiSal
17000.00	105	0.98	HiSal
17000.00	105	0.98	HiSal
.....			
9500.00	85	0.79	HiSal
9500.00	85	0.79	HiSal
9500.00	85	0.79	HiSal
9000.00	81	0.76	Unknown
9000.00	81	0.76	Unknown
.....			
7900.00	71	0.66	Unknown
7800.00	70	0.65	Unknown
7700.00	69	0.64	LowSal
7500.00	67	0.63	LowSal
7500.00	67	0.63	LowSal
.....			
2200.00	2	0.02	LowSal
2200.00	2	0.02	LowSal
2100.00	1	0.01	LowSal

Fig.7: constructing a non-stationary fuzzy query statement and the result of executing it.

```
select Count(Salary) ,
SallngVal(Salary,Percentage) as Degree
from employees_salary_view
Group By SallngVal(Salary,Percentage)
```

Run Query

COUNT(SALARY)	DEGREE
23	HiSal
15	Unknown
69	LowSal

Fig.8: results of execute a non-stationary fuzzy query that asks about the count of employees according to each category of salary

5.3 FSQL Builder Module

This module is intended to construct Fuzzy SQL statements through a simple graphical user interface as shown in Fig.9. Accordingly, the user can choose what he wants to write rather than the sentence typed by the user to help deal with the system.

The screenshot shows a 'Query Builder' window. On the left, there are checkboxes for 'Employee', 'Location', and 'Dept'. On the right, there are checkboxes for 'Employee Empno', 'Employee Empname', 'Employee Job', 'Employee Dept', 'Dept Dept', and 'Dept Dname'. A 'Build Select Statement' button is in the center. Below it, a text box displays the generated SQL query:


```

    Select
    Employee.Empno,Employee.Empname,Employee.Job,Dept.Dept,Dept.Dname
    From Employee,Dept
    Where Employee.Dept=Dept.Dept And (Dept.Dept = 1 OR Dept.Dept = 2)
    Order by Dept.Dname DESC
    
```

 Below the query, an 'Execute Select Statement' button is visible. Underneath, a table shows the results of the query:

Empno	Empname	Job	Dept	Dname
2	Retai Aymam	Web Developer	2	Information System
1	Dr Ahmed	Prof	2	Information System
1	Ali Maher	Engineer	1	Computer Science
3	Dr.Hisham	Chair Man	1	Computer Science

 At the bottom of the window, there is a table for defining column aliases, sort types, filters, and functions:

Column Name	Alias	Sort type	Filter	OR	Function	Having
Employee Empno		None			None	
Employee Empname		None			None	
Employee Job		None			None	
Dept Dept		None	= 1	= 2	None	
Dept Dname		Ascending			None	

Fig.9: Results of execute non-stationary fuzzy query

6 Conclusion

In this paper a new technique for applying non-stationary fuzzy set with fuzzy query of relational databases has been introduced. Non-stationary fuzzy set is a fuzzy set with dynamic definition. In other words the definition of non-stationary fuzzy set is sensitive to some factors or variables in the related environment accordingly, using non-stationary fuzzy set make fuzzy query of relational databases more flexible and sensitive to the current database state. The result reflects the nature of the data existing in the current database state. An illustrative example was introduced to show the benefit of using non-stationary fuzzy sets with fuzzy query.

7 References

- [1] Mishra J., "Fuzzy Query Processing", International Journal of Research and Reviews in Next Generation Networks, Vol. 1, No. 1, March 2011.
- [2] Grissa A. and Ben Hassine M., "New Architecture of Fuzzy Database Management Systems", The International Arab Journal of Information Technology, Vol. 6, No. 3, July 2009.
- [3] Qi Yang et al., "Efficient Processing of Nested Fuzzy SQL Queries in a Fuzzy Database", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 13, NO. 6, NOVEMBER/DECEMBER 2001.
- [4] Abbaci K. et al., "Selecting and Ranking Business Processes with Preferences: An Approach Based on Fuzzy Sets", International Conference on Cooperative Information Systems, Greece (2011)
- [5] Singh K. et al., "Study of Imperfect Information Representation and FSQL processing", International Journal of Scientific & Engineering Research Volume 3, Issue 5, May-2012
- [6] Wibisono O., "Simple Fuzzy SELECT Query on Crisp Database", MakalahIF4058 TopikKhususInformatikaI – Sem. II Tahun 2011/2012
- [7] Garg A., and Rishi R., "Querying Capability Enhancement in Database Using Fuzzy Logic", Global Journal of Computer Science and Technology, Vol. 12 Issue 6 Version 1.0 March 2012.
- [8] Jain N. and Shukla S., "Fuzzy Databases Using Extended Fuzzy C-Means Clustering", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, May-Jun 2012.
- [9] Olufade F. and Oladeji, P., "A FUZZY SEARCH MODEL FOR DEALING WITH RETRIEVAL ISSUES IN SOME CLASSES OF DIRTY DATA", Proceedings of the 16th International Conference on Information Quality, 2011

- [10]Ashu, "Intuitionistic Fuzzy Approach to handle Imprecise Humanistic Queries in Databases", International Journal of Computer Applications (0975 – 8887), Volume 43– No.20, April 2012.
- [11]Ravi G.,SANTHOSH J., "A survey on retrieving Contextual User Profiles from Search Engine Repository", International Journal Of Computational Engineering Research, Vol. 2 Issue. 4, 2012.
- [12]Jonathan M. Garibaldi, MarcinJaroszewski, and SalangMusikasuwan., "Nonstationary Fuzzy Sets", IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 16, NO. 4, AUGUST 2008.
- [13]Klir & Yuan. Fuzzy Logic & Fuzzy sets.
- [14]<http://www.dmitry-kazakov.de>, 1/12/2012
- [15]GadAllah A. and Hefny H. "An Efficient Database Query Processing Tool Based on Fuzzy Logic", The 37th annual conference on statistics, 2002.
- [16]Shawky A. et al. "FRDBM: A Tool For Building Fuzzy Relational Databases", The Egyptian Computer Journal, 2006.

Access Control Models for Vector-based Data

Mennatallah H. Ibrahim¹ Nermin Hamza² Mohamed Nour El-Deen³ Hesham A. Hefny⁴

Abstract

The research on spatial database is mainly concentrated on spatial data modeling, spatial query languages, spatial indexing, query processing and optimization, spatial data mining, etc. In practice, the problem of spatial information security is more important. However, spatial data security has not been much investigated yet by the Geographical Information System (GIS) community. Current applications increasingly require spatial data to be managed remotely either through the Web, through local networks or even through mobile devices. The users of these applications do not need only to display the data but also to interactively modify it. As a result spatial data is exposed to various security risks as it may be displayed to or altered by illegitimate users.

Various access control models have been proposed for conventional databases. However, these models are not appropriate for geographical databases, due to the special nature of spatial data. Fortunately, several access control models that are built especially for spatial data either vector- based or raster-based have been proposed. This paper provides a brief overview of the work done in the area of vector-based access control models and proposes a framework for a secured server-side web-based GIS.

Keywords

GIS, Access control, vector-based data, Authorization

1. Introduction

The proliferation of geographical information services on the web, through local networks or even through mobile applications is creating great opportunities for the shared use of spatial data. Spatial data may contain sensitive information, for this reason data must not be freely disclosed 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. This can be achieved by enforcing controlled access to the spatial data.

The confidentiality and integrity of spatial data can be ensured by enforcing access control to it. 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]. In general, when an access control system is applied, data access is regulated by an access control policy. Basically, data access is controlled through a set of authorization rules which states who can access which resource for doing what.

^{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

Each authorization rule, in its basic form, consists of a triple <subject, object, privilege>. The subject indicates who can access the data resources. The object is a resource identifier. The privilege is the kind of action that can be performed by the subject on the given object. Such rules are defined in accordance with an access control model. When a user tries to access a resource, the request is checked against the set of authorization rules. If the match is successful the user can access the resource; otherwise the request is discarded. Authorization rules are granted in accordance to an administration policy [2].

Spatial data format can be classified into raster and vector. Raster and vector are different but common data formats used to store spatial data. Vector data use X and Y coordinates to define the locations of points, lines, and polygons that correspond to map, while raster data, use a matrix of square areas to define where features are located. These squares are called pixel. Pixels are typically of uniform size, in which their size determines the details that can be maintained in the dataset. Because raster data represent square areas, they describe interiors rather than boundaries this is the same case with vector data. Vector data are excellent for capturing and storing spatial details, while raster data are well-suited for capturing, storing and analyzing data that vary continuously from location to another, such as elevation, and temperature. Raster data formats also are used to store aerial photography and satellite imagery [10].

For a long time, work on access control models for spatial data has mainly dealt with raster maps, focusing on protecting sensible information that can be revealed by high resolution image satellites. In such model, each protected object is basically an image or portion of an image. Thus, the model supports neither the vector-based representation of features nor the topological one. As such, the model is not adequate for usage in current GIS and spatial database management system. Moreover, it only deals with read privileges and it is thus not adequate for dynamic applications that require data modifications [3]. The first access control model that dealt with vector-based data and topological representation of features was proposed in 2004.

The paper is organized as the following: The paper is divided into five sections: Section 2 presents an overview of web vector-based access control models. Section 3 presents an overview of local networks vector-based access control models. In section 4, we propose a framework for a secured server-side web-based Geographic Information System. The work is summarized in the last section.

2. Web Vector-Based Access Control Models

Vector-based access control models can be classified into web, mobile [4] and local networks vector-based access control models. Web vector-based access control models can be further classified into fine grained and not fine grained models (see Figure 1). Fine grained access models allow administrator to specify authorizations against map objects at a very fine granularity level, taking into account the various spatial representations, the object dimension and its non spatial characteristics. For example, the administrator can authorize the user to see a given object at a specific dimension and not at higher dimensions, thus hiding detailed information about the object shape, while not fine grained access control models allow authorizations to user depending on the user's role against all allowed map objects regardless their spatial representation or their dimensions.

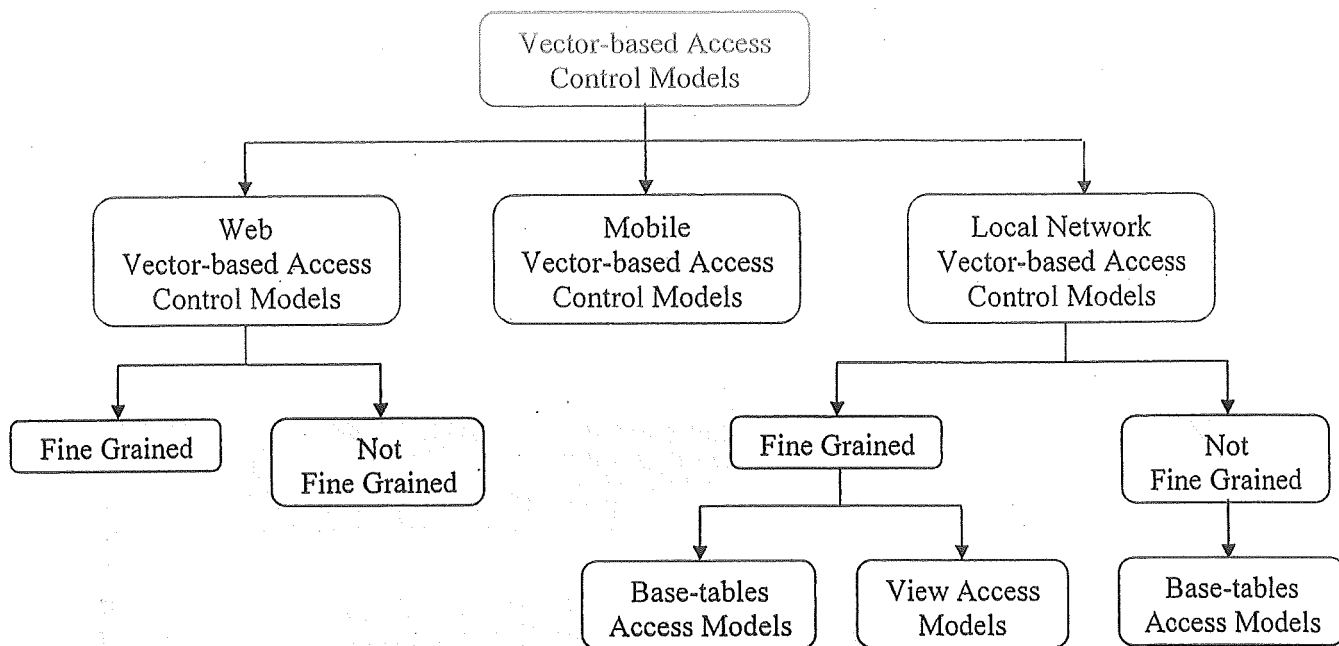


Figure 1: Vector-based Access Control Models Classification.

The first access control model for vector-based data was proposed in 2004 [1] by E. Bertino and M. Damiani, in which its goal is, protecting vector-based data on the web. This model is based on an extended notion of authorization rule, where it's 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 the proposed model, an authorization rule is defined as a tuple consisting of six fields [1]:

- (i) Role: the role 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;
- (v) Grantor: the grantor 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.

The second model [2] was proposed as well in 2004 by M. Damiani and D. Momini. This model is similar to the first proposed model, where it proposes an extension to the classical access control model, which is based on the notion of authorization rule, in which its goal is, protecting vector-based data on the web. The central idea is assigning a geographical scope to authorization, namely a bounded region in which the authorization is valid. In the proposed model, an authorization rule is defined as, a tuple consisting of the same six fields as the first model and data were manipulated through the operations provided by a Web Map Management Service that mediates between users and data.

The advantage of Bertino and Damiani models is represented in, being the pioneer models which provides access control to spatial data, in order to ensure it's confidentiality and integrity so they presented the base of many other models. While their disadvantages are represented in, not allowing one to define authorization rules for objects at a finer level of granularity and allowing all users to access the base tables of the spatial database regardless of their roles.

In 2004 [3], a model that addresses the limitation of Bertino and Damiani models was proposed by A. Belussi and E. Bertino where it allows the security administrators to specify authorizations against map objects at a very fine granularity level on the web.

Belussi model is different from Bertino and Damiani models in the following [3]: (a) the map model is more complex. (b) Providing multiple representations (geometric and topological) of the same map object and multiple representations of the same feature. (c) The previous models propagation is only allowed along application-dependent privilege hierarchies, on the other hand, in this model propagation rules constituted an invariant part of the access control model and are defined along object and privilege hierarchies. Similarly to the previous models, it's goal is protecting vector-based data on web and it supports the concepts of authorization window.

In this model two new concepts are supported, 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. An authorization is a tuple which contains subject, object, privilege, window, authorization sign, grantor, grant_option and authorization type.

The advantages of Belussi model 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 model efficiency. While it's disadvantage is represented in, allowing all users to access the base tables of the spatial database regardless of their roles.

By comparing the three models, it is clear that the three presented models support the concept of authorization window and that the Belussi model is more efficient than Bertino and Damiani as it addresses the limitation of the lack of fine granularity and it provides more protection to spatial data through fine granularity, authorization type and authorization sign which are not available in the other models. The differences between the three models are represented in Table 1.

Table 1: Web Models Comparison

Models	Fine Granularity	Authorization Window	Authorization Sign	Authorization Type
Bertino Model [1]	—	✓	—	—
Damiani Model [2]	—	✓	—	—
Belussi Model [3]	✓	✓	✓	✓

3. Local Networks Vector-Based Access Control Models

Local Networks Vector-based Access Control Models can be classified into fine grained and not fine grained models and further into base-tables access and view access models as shown in Figure 1. Base-table access control models allow all users to access the base tables of the spatial database regardless of their role, while the view access models allow only certain roles (i.e: administrators) to access base tables, while any other user access only views of the base tables but not the base tables themselves.

In 2007 [5], Y. Zeng and Z. Wei proposed an access control model in order to overcome the shortcoming of Security Socket Layer (SSL) protocol. The model also presents a new concept, which is certificate and private key cruising. They introduced Public-Key Infrastructure (PKI) and SSL protocol to resolve security problems of spatial database, with which identities were authenticated using digital certificates before data transmission and data are encrypted during it's transmission.

SSL has a fatal shortcoming, which is the absence of the authority management to different roles during identity authentication [5]. This model presents a new solution for this shortcoming, where access control mechanism is introduced into spatial database without additional security model nor a structure change of certificates. Combining with Role Based Access Control, a role control table on the spatial database server is built, where the relationship of roles and authorities is saved. When users access the spatial database, they transmit their own certificates and roles to the database server, and then the server confirms the validity of their identities based on their certificates and lookups the role control table based on their roles in order to decide their authorities.

The model's framework core part is, Certificate Authority Service (CAS), its main task is to publish certificates of the server agent, clients and Registration Authority (RA) and manage these certificates. RA manages registration of all clients and distributes roles to them. The server agent is an agency between Spatial Database Engine (SDE) and spatial database, with which clients submit their access requests to Spatial Database Management System (SDBMS) and SDE transmits returns from the server to clients.

The advantage of Zeng model is, presenting a new concept which is certificate and private key cruising that ensures more protection to the spatial data. While it's disadvantages are, the model neither supplies fine grained access control nor supports authorization window and it allows all users to access the spatial database base tables regardless of their roles.

In 2008 [6], J. Lin and Y. Fang proposed a view-based authorization model, which exploits the built-in mechanisms of spatial database to realize fine-grained spatial access control. From it's advantages, there is no need to modify software, it is easy to implement, and it has low potential to leak information. Restriction could be performed on columns and rows as well.

After views satisfying access control requirements are created, the next thing is to grant access rights for specific users to intended views. The procedure includes two steps; the first step is to prevent users from direct access to base tables. The second step is to grant access rights for views to intended users.

The view-based access control reference framework has the following components [6]: user applications, database user accounts, privileges, views, Base tables, and a policy enforcement module. The views (View1, View2 ...) are created from the same base table in spatial database according to different access control requirements, either spatial, non-spatial or their combination, and then grant them to different database user accounts with corresponding access privileges. The policy enforcement module is responsible for assigning a proper database user account to the user application based on some criteria, for example positions and responsibilities in an agency or organization.

The advantages of Lin model are represented in, supporting fine granularity, not allowing users to access base tables themselves but only views of them, views do not duplicate the table data and thus saves a large amount of storage space, combination of the two kinds of view restrictions, non-spatial and spatial access controls can be seamlessly integrated into a single definition of the view, and there is no need for view-based model to alter the schemas of base tables or add extra tables to store authorization information. While it's disadvantage is the concurrent control where although the mode of multi views to a single base table saves storage space and overhead of maintaining consistence of multi data replications, it will cause the problem of concurrent control, especially when many users are accessing those views based on the same base table concurrently.

In 2009 [7], M. Yan and Y. Gao proposed vector-based access control model for grid environment. This model supports the concept of fine granularity. In this model, every user is mapped to a role, which is a collection of resources and can be defined as a tuple: $R1 \langle \text{Rolename}, \text{Resource}, \text{Permission} \rangle$. In this tuple, resource is defined as a set of spatial datasets and services with spatial control: $R2 \langle \text{ResourceID}, \text{Rectangle} \rangle$. In tuple $R2$, ResouceID is the unique identification of geospatial data or related services Rectangle is Maximum Bounding Box (MBR) of related resources. Then based on the tuple $R1$ and $R2$, a common grid user can be defined as $R3 \langle \text{LoginID}, \text{Password}, \text{Rolename} \rangle$, where role name is the related role of user that is decided by administrator. The role is classified as two types; administrative roles and common roles. All common users are mapped to common roles, while the administrator users are mapped to administrative roles.

The advantages of Yan mode is represented in, supporting fine granularity concept. While its disadvantage is, allowing all users regardless of their roles to access the base tables.

In 2010 [8], F. Ma and Y. Gao proposed a fine grained access control model which based on Role-based Access Control model for grid environment. In this model, the authorization mechanism is a double authorization, which is gradually refined. Similarly to RBAC, by the first authorization, which is the user-role maps, users get the appropriate permissions of the role. It is the authorization to layers, with a horizontal layer as a unit. The Secondary authorization judge whether the user has access to the data within a region based on the user-related attribute information. It is processed through layers in the vertical direction.

The first authorization is similar to the RBAC model in authorizing the role, and the second authorization authorizes the specific user based on his attribution. The role authorization uses the most common method, the classic triple tuple, whose structure is: $\langle R, RS, M \rangle$, where, R represents an authorized role, RS represent the resources that the role has access to, and M represents the access model, including the Read or Write. The second authorization is used to determine whether a specific user has the right to access to the specific region dynamically and then authorize the user, based on the user attribute. The authorization structure is shown as follows: $\langle U, P, T \rangle$, where, U refers to the authorized user, P is the polygon information of the specific region needed to authorize, and T refers to the type of authorization, which can be positive or negative.

The advantages of Ma model are, making the fine-grained access control come true, where the user can only have an access to parts of the layers and supporting the concept of authorization sign. While its disadvantages are not supporting view access concept, and also an authorization conflict between the first and the second authorization may occur.

In 2012 [9], K. Tuyen Le Thi and T. Khanh Dang proposed an access control model called Spatial Temporal Role Based Access Control (STRoBAC), in which their objective is to propose a model that can support spatio-temporal aspects and other contextual conditions, in addition to access control that based on the role of subject. Unlike the previous models that are based on the characteristics of geospatial objects or subjects (e.g. feature type, specific area of object, location of requester, his/her roles within the valid area, etc.) only, STRoBAC provides the specification and enforcement of security policies supporting fully the kinds of restrictions [9]. Also unlike other models, beside the context holds between subject, action and object, STRoBAC proposed other additional contexts (e.g. contexts of assigning subject in role, using object in view, and assigning action in activity).

STRoBAC based on Spatial Temporal Role Based Access Control (STRBAC) model, in order to describe location and time information; express the contextual conditions based on Organization Based Access Control (Or-BAC) model. Similar to Or-BAC model, STRoBAC had basic sets of entities: S (subjects), A (actions), O(objects), R (roles), A (activities), V (views) and C (contexts).

The advantages of Tuyen model are, supporting the following, the concept of fine granularity, the spatio-temporal aspects and other contextual conditions, in addition to access control that based on the role of subject. Besides being view-based access control model that prevents users from accessing base tables.

By discussing the various local network models, it is clear that most of the models support the concept of fine granularity. Each model has a specific capability or feature that improves the access control to local networks spatial data. Unlike most of the local network models, Zeng model does not support fine granularity but it concentrates more on overcoming the shortcoming of SSL protocol and using PKI and digital certificates in order to provide a model that efficiently control access to contained spatial data. Otherwise, Lin Model supports fine granularity and also supports the concept of view-based access, which is a solution for protecting the access to base tables of the spatial database by preventing users from accessing the base tables themselves but only allowing them to access views from these tables according to their roles. Yan Model was a fine grained model that supports the concept of authorization window in grid environment. Ma Model is a fine grained access control model; it is based on double authorization in order to control the access to the contained spatial data (authentication to layer and authentication to specific region of that layer). Tuyen model is a fine grained model that supports spatio-temporal aspects and other contextual conditions in addition to, access control that based on the role of subject. A brief comparison between the proposed local network access control models is shown in Table 2.

Table 2: Local Network Access Models Comparison

Models	Fine Granularity	Time-based	Position-based	Authorization Window	Authorization Sign
Zeng Model [5]					
Lin Model [6]	✓				
Yan Model [7]	✓			✓	
Ma Model [8]	✓			✓	✓
Tuyen Model [9]	✓	✓	✓		

4. Proposed Framework

We propose a framework for a secured server-side web-based GIS. The proposed framework is based on the access control models which are proposed in [1], [3] and [6]. The goal of the proposed framework is to ensure security of spatial data on the web. The proposed framework is illustrated in Figure 2.

The proposed framework is based on the well-known three-tier architecture which consists of; presentation, application and data storage layers. First, the presentation layer, which displays the output of the application layer to the user and also user's requests are posted to the application layer through the presentation layer.

Second, the application layer, which forms the centre of the framework; it is responsible for overseeing the relationship between all components in the system, and determining how the system functions. The application layer consists of two main services; 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. It has to be noticed that the application layer also includes an Authentication Service that can be based on username/password or provide more complex service. The application layer is also responsible for performing modifications to the database layer that are requested by the client.

Third, the data layer, which is responsible for the storage and the provision of data to the application layer. Databases employ Database Management Software (DBMS) that handle a number of issues including concurrent user access, and the efficient storage of data.

The proposed framework follows the server-side approach, where the GIS software resides on the server and the browser acts as a GUI at the user end. The user accesses GIS functionalities through the browser. The user requests or initiates a function over the Internet to the server, which then activates GIS software which in turn processes the request and sends results back to the user via the browser. The GIS software starts and runs the analysis requested by the user. The output data from GIS software returns to the browser for presentation to the user.

The proposed framework has the following components: user applications, database user accounts, privileges, views, Base tables, and a policy enforcement module. The views (View1, View2 ...) are created from the same base table in spatial database according to different access control requirements, either spatial, non-spatial or their combination, and then grant them to different database user accounts with corresponding access privileges. The policy enforcement module is

responsible for assigning a proper database user account to the user application. After views satisfying access control requirements are created, the next thing is to grant access rights for users to intended 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. Secondly, a role and a database user account will be assigned to the authenticated user. Thirdly, the user will be allowed to issue request(s) to the system. Each request from the user will be mapped onto one or more operations of the Application Service. Application Service in turn will interact with the Access Control Service to verify whether the user request can be performed or not according to his role and database user account. If the request can be performed, the Application Service will extract the data requested by the user from the data layer and then process it and pass the result to the presentation layer to be displayed to the user.

The framework also grants specified access privileges to administrator users. In principle, we cannot modify base tables through their views. Therefore, as for such privileges as INSERT, UPDATE, and DELETE, there is an additional step to create rules to implement their modification functionality.

The advantages of this framework represented in providing fine granularity access control to spatial data and supporting the concepts of authorization sign and authorization type which will provide more efficient and secure access control to the spatial data. Also the framework will prevent any user from accessing the base tables in order to protect them from being altered by or disclosed to illegitimate users, instead, users will be allowed to access views from these base tables.

The differences between the proposed framework and the previous web-based GIS frameworks are; this framework prevents users from accessing the base tables and allows accessing only views from those base tables to be accessed, and it will assign to the users not only roles but also database user accounts once they are authenticated to the system that determine which views the user can access according to their roles and this provide more efficiency to the access control process. Other differences are shown in Table 3.

Table 3: Proposed and Previous Models Comparison

Models	Fine Granularity	Authorization Window	Authorization Sign	Authorization Type	Views Access
Bertino Model		✓			
Damiani Model		✓			
Belussi Model	✓	✓	✓	✓	
Proposed Model	✓	✓	✓	✓	✓

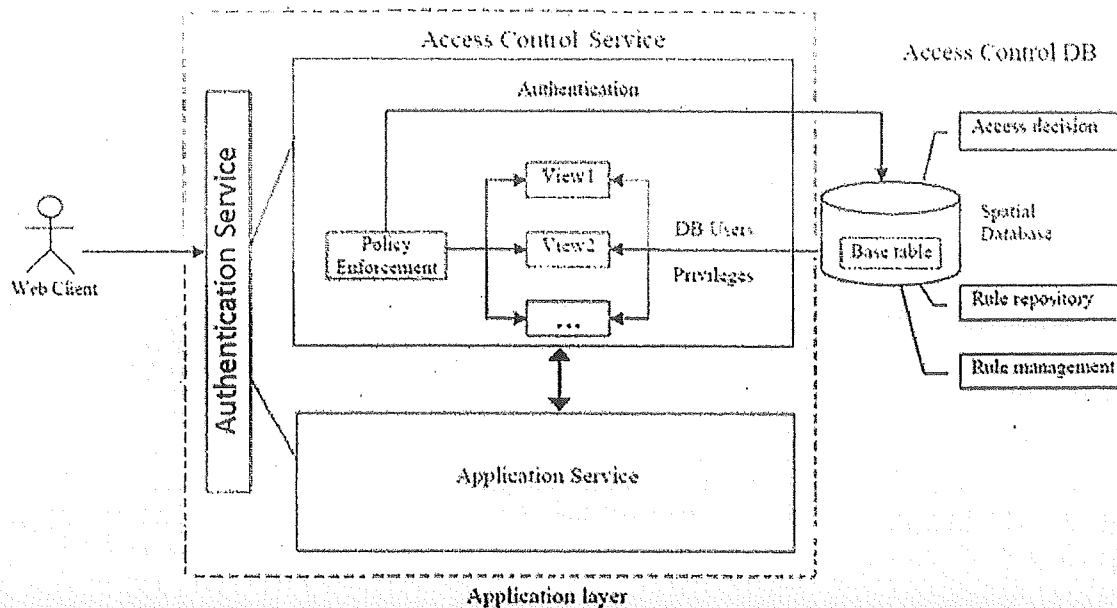


Figure 2: Framework for secured server-side web-based GIS.

6. Conclusion and Future Work

Access control is very critical and important for spatial data in order to keep its confidentiality and integrity. Spatial Data requires specific access control models due to its special nature. Several access control models have been proposed for both vector-based and raster based-data. For a long time proposed models dealt only with raster data. In 2004 the first access control model for vector data was proposed. Access control models for vector-based data can be classified into web, mobile and local networks access control models. The web access control model can be further classified into fine grained models and not fine grained models. Also the local networks access control models can be classified into fine grained and not fine grained models and further classified into base-tables access and view access models.

In this paper, we proposed a framework for secured web-based server-side GIS based on three of the presented access control models. 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.

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.
- [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.
- [3] A. Belussi, E. Bertino & B. Catania. "*An Authorization Model for Geographical Maps*". Proceedings of the 12th annual ACM international workshop on Geographic information systems. 2004.
- [4] E. Bertino, B. Catania & M. L. Damiani. "*GEO-RBAC: A Spatially Aware RBAC*". Journal of ACM Transactions on Information and System Security. 2005.
- [5] Y. Zeng, Z. Wei, Q. Yin. "*Research on Spatial Database: A Secure Access Mechanism*". Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong. 2007.
- [6] J. LIN, Y. FANG, B. CHEN & L. WAN. "*View-based Access Control Mechanism for Spatial Database*". International Conference on Earth Observation Data Processing and Analysis (ICEODPA). 2008.
- [7] M. Yan, Y. Gao, L. Wu, and others. "*Spatial Data Access Control in Grid Environment*". Geoinformatics, 2009 17th International Conference. 2009.
- [8] F. Ma, Y. Gao, M. Yan and others. The Fine-Grained Security Access Control of Spatial Data. Geoinformatics, 2010 18th International Conference .2010.
- [9] K. T. Le Thi, T. K. Dang, P. Kuonen & H. C. Drissi, "*STRoBAC – Spatial Temporal Role Based Access Control*". 4th International Conference on Computational Collective Intelligence (ICCCI 2012). 2012.
- [10] <http://www.extension.org/pages/17809/understanding-the-difference-between-raster-data-and-vector-data>

An Enhanced Model For Linguistic-based Fuzzy Ontology

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

Abstract

Uncertain and imprecise information exist in many real-world applications and for this reason, fuzzy ontologies have been extensively investigated and increasingly created. Recently, researchers concentrate their work about how to create and store very large fuzzy ontology in an efficient way. Regarding its creation, fuzzy ontology can represent multi domains in a certain view. The relation between each domain or each fuzzy ontology component is represented as a membership degree. To represent a certain fuzzy domain in more than one view, a fuzzy ontology is created for each of these views. Concerning storing fuzzy ontology, researchers tend to store it in databases to decrease query response time. This paper proposes a general and flexible model that can build a large multi-domain fuzzy ontology in multi-view. The relation between each domain or each fuzzy ontology component can be represented as fuzzy numbers or linguistic terms. The model stores it in a relational database.

Keywords

Fuzzy ontology, fuzzy ontology components, fuzzy ontology storing

1. Introduction

During the last decade, Ontology have seen a great development, as the result of the widely use of knowledge-base systems. 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 [5]. When talking about ontology, many aspects should be considered. According to our model, we will concern about two of them. These two concepts are: ontology representation, and ontology storing techniques.

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 domains. 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 theory into ontology [21].

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

Then we can generate fuzzy ontologies, which contain fuzzy concepts and fuzzy memberships. Fuzzy ontologies are capable of dealing with fuzzy knowledge, and are efficient in text and multimedia object representation and retrieval [4]. The rest of this paper is organized as follow section 2 and section 3 present an overview on fuzzy sets and fuzzy ontology respectively. Section 4 shows some related works. Our proposed model is discussed in section 5. Section 6 supports our model with a case study. Finally, conclusion and future work is presented in section 7.

2. Fuzzy Sets

The value that any linguistic variable take can have different forms. These forms are crisp value, fuzzy number, linguistic value, quantified (hedged) linguistic value, complex value, see table1 for an example. As table1 shows, crisp value is a numerical value. Fuzzy number is an approximated value such as, about 30, approximately 30, around 30... etc. Linguistic value is a linguistic term that describes the value of a linguistic variable in natural language terms. Hedged linguistic value uses a hedge such as, very, extremely, with linguistic terms. Complex term is a combination of hedges, linguistic value, and fuzzy operators, such as NOT, AND, OR.

Table1: shows the different values that can be assigned to the linguistic variable temperature.

Linguistic Variable	Value				
	Crisp value	Fuzzy value	Linguistic value	Quantified (hedged) linguistic value	Complex value
x='temperature'	30	Around 30	Cold, hot	Very cold	Not very hot

Fuzzy numbers and linguistic values are represented using fuzzy sets. Fuzzy set is a set whose elements belong to it with a certain membership degree for each element; also such elements can belong to any other set with a certain membership degree. It is written in discrete domain as: $A = \{\sum_1^n \frac{\mu_A(x_i)}{x_i}\}$, while in continuous one: $A = \{\int \frac{\mu_A(x)}{x}\}$, where A is the fuzzy set, n is the number of elements in this set, x_i is an element in this set, μ is the membership value that represents to what extent does x_i belongs to the set A. Membership functions are used for defining fuzzy sets mathematically. Fig. 1 shows a triangular membership function representing the fuzzy value around(30), which can be expressed over discrete domain as: around 30= {0/25, 0.1/26, 0.5/27, 0.6/28, 0.9/29, 1/30, 0.9/31, 0.6/32, 0.5/33, 0.1/34, 0/35}. Fig.2 shows two triangular fuzzy sets representing the linguistic values 'cold' and 'hot' that can be expressed over a discrete domain of temperature values as: Hot={.....,0.5/30, 0.6/37, 1/40...}, Cold= {..., 0.5/6, 0.6/7, 1/10 ...}.

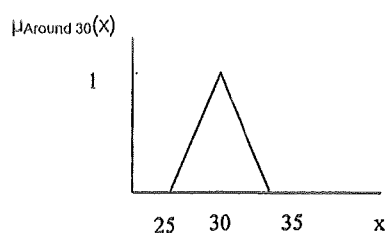


Fig. 1: Fuzzy set representing the fuzzy number "around 30"

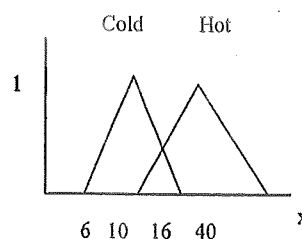


Fig. 2: Fuzzy sets representation of two linguistic values "hot" and "cold"

Table 2 shows some of the most used membership functions. Similar to crisp sets, fuzzy sets have three basic operations. These operations are: union, intersection, and complement. Union operator, also known as disjunction or 'OR', can be defined via its membership function as:

$$= S(\mu_A(x), \mu_B(x))$$

where S is s-norm. Intersection operator, also known as conjunction or 'AND', can be defined via its membership function as:

$$= T(\mu_A(x), \mu_B(x))$$

where T is t-norm. Complement operator, also known as NOT, is defined via its membership function as:

Table 2: the most well-known membership functions

Membership Function Form	Mathematical Formula
Triangular 	$\text{Triangle}(x:a, b, c) = \max(\min(\frac{x-a}{b-a}, \frac{c-x}{c-b}), 0)$
Trapezoid 	$\text{Trapezoid}(x:a, b, c, d) = \max(\min(\frac{x-a}{b-a}, 1, \frac{d-x}{d-c}), 0)$

Hedges are adverbs, or adjectives. They are used to modify the membership function of the used linguistic value using the mathematical power operator.

To convert fuzzy value to crisp one, defuzzification technique is used. There are several definitions for defuzzification. The most popular one is the Center Of Area technique (COA). COA is defined as:

$$\text{COA}(A) = \frac{\sum_x \mu_A(x)x}{\sum_x \mu_A(x)}$$

3. Fuzzy Ontology

Since 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 [10]. Researchers define fuzzy ontology components according to their used application and domain. Following we will show some of these definitions:

- 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 [11].
- Fuzzy ontology is a quadruple(C, R, F, U), where C is a set of concepts, R is a set of fuzzy abstract relations, F is a set of fuzzy concrete relations, and U is the universe of discourse [17; 18].
- Fuzzy ontology is a quadruple (C, FC, R, A), where C is a set of concepts, FC is a collection of sets of fuzzy attributes (one set for each concept), R is a set of relations (including taxonomic relations), and A is a set of axioms [12].
- 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 [13; 14].
- 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 [15; 16].

Commonly, in literature, there are many aspects concerned with fuzzy ontology such as creation, storing and inferring. The proposed fuzzy ontology model presented in the paper emphasizes two aspects namely fuzzy ontology representation and storing in databases. Representing fuzzy ontology means representing its components either through the fuzzy extension of OWL or through the fuzzy extension of DL, Fuzzy Description Logic (f-DLs). Both of them are capable of representing vague and imprecise knowledge using fuzzy set theory. Fuzzy Description Logic (f-DLs) is the extension of the classical description logic. F-SHIN (D) and f-SHOIN (D) are the fuzzy extension for SHIF (D) and SHOIN (D) respectively. Similar to DL, f-DL consists of a knowledge base (KB) and some reasoning capabilities. It has three types of elements: concepts, roles, and individuals. Concepts denote fuzzy sets of individuals. Roles denote fuzzy binary relations. A KB comprises different types of knowledge, grouped into some different components:

- Extensive knowledge is a particular knowledge about some specific situation. This type of knowledge is contingent (it depends on several circumstances and is subject to changes). It is stored in an Assertional Box (ABox).
- Intensive knowledge is a general knowledge about the application domain, which is usually unchangeable. This type of knowledge is structured in a Terminological Box (TBox) with information about the concepts, and a Role Box (RBox) with information about the roles.

Storing fuzzy ontology means storing fuzzy ontology structure and fuzzy ontology individuals. In general, fuzzy ontology can be stored either in files or in databases. Compared with database-based storing technique, files greatly reduces the load and update time. However, database systems provide many query optimization features, thereby contributing positively to query response time. Since fuzzy ontology reasoner is used to answer user's query over huge amount of knowledge, database will be the best method for storing it [6]. But, unfortunately a little work has been done in this area.

4. Related Work

The work stated in [7] is proposed to store fuzzy ontology (structure and instance) in a relational database. It assumes that the structural components of fuzzy ontology include only concepts, properties, and fuzzy relation between concepts. These fuzzy relations are represented using membership degrees. Another model in [20] called Fuzzy Relational Ontological Model (FORM). It uses fuzzy ontology in information retrieval system for expanding user's query. FORM adds a new ontology constructor, term constructor, so that its ontology components becomes; concepts, terms, and fuzzy relations between concepts and terms. These fuzzy relations are represented using membership degrees. [19] uses fuzzy ontologies in information retrieval system. It works on fuzzy ontologies that are related to each other with membership degrees.

The work presented in [9] stores only crisp ontology in relational database considering concepts, properties, and relations as the main ontology components. Then, it uses pl/sql procedure to build an inference engine. [3] Extracts ontologies from OWL files and then stores them in relational databases. [1] Proposes a schema to store ontologies with fuzzy datatypes into a database. This database schema allows the storage of ontology along with its instances preserving all information. Ontology and instances are stored in different schemas in order to improve the access to instances while retaining the capacity of reasoning over the ontology.

5. The proposed model

5.1 Fuzzy ontology definition

The proposed model is a Multi Views-Fuzzy Related Ontologies. Its idea is inspired from [19, 20], where [19] works on multi fuzzy ontologies, while [20] assumes fuzzy ontology as a set of concepts, terms, and relation between them. The proposed model is a two or more fuzzy ontologies related with each other with fuzzy numbers. Relations in this model can have one or more values, each reflects a certain view. Its main features can be summarized as follow:

- It is a more flexible and generic fuzzy ontology, which can cope with any domain and any application.
- The proposed model main components are concepts, relations, properties, terms, and individuals.
- Having multi-fuzzy-values representing multi-point of view as the relations between the fuzzy ontology components or the related fuzzy ontologies, e.g., In the old English, poetry represents

the English literature with degree about 0.3, while in the modern English, poetry represents about 0.25 from the English literature.

- Using linguistic values and fuzzy numbers to express the relations between fuzzy ontology components.
- Using linguistic values and fuzzy numbers to express the relation between the related fuzzy ontologies.
- The used linguistic values and fuzzy numbers are defined by the domain expert according to his/her own subjective view.

Each fuzzy ontology in our proposed model is a couple of (f-OS, f-OI), where

- f-OS is a fuzzy ontology structure (TBox and RBox),
- f-OI is a fuzzy ontology individual of concepts and relationships associated with the fuzzy ontology structure (ABox).

F-OS is defined as a tuple f-OS (C, CR, P, T, A), where

- C is the domain concept or an ontology concept. C has the same definition in f-DL-SHIN.
- CR is a relation between the concept C and another concept in the domain. CR is defined as the tuple CR (R, C2, FN1, FN2, ..., FNn) where
 - R is the relation name that concept C has,
 - C2 is a concept, that C is related with,
 - FN1 is a fuzzy number expresses the degree to which C is related to C2 using the relation R in a certain view.
- P is a concept property. Property can be object property (OP) or datatype property (DP). Op can be defined as the tuple OP=(OP, C2, FN1, FN2, ...FNn, OPR), where
 - OP is the object property name,
 - C2 is a concept; it is the value of the object property,
 - FN1 is a fuzzy number that expresses the degree to what extent concept C is related to the concept C2 through the property OP in a certain view.
 - OPR is the relation between the object property OP and another one. OPR is defined as the tuple OPR (OPR, OP2, FN1, ... FNn), where
 - OPR is the relation name that the object property OP has,
 - OP2 is an object property, that OP is related with,
 - FN1 is a fuzzy number that expresses the degree to which OP is related to OP2 with the relation OPR in a certain view.
- T is a fuzzy set of terms that define the concept C. T is defined as the tuple T (T, FN1, ..., FNn, TR), where
 - T is the term name,
 - FN1 is a fuzzy number that expresses the degree to which T can express the concept C according to a certain view,

- TR is a term relation with other term. TR is defined as the tuple TR (TR, T2, FN1, ..., FNn), where
 - TR is the term relation name that the term T has,
 - T2 is a term, that T is related with,
 - FN1 is a fuzzy number that expresses the degree to which the term T is related to T2 with the relation TR in a certain view.

5.2. Storing the proposed model in relational database

This section presents our contribution in storing the proposed model in relational database. It can be summarized as follow:

- Storing all ontology components after stemming it using porter stemming algorithm.
- Sorting different point-of-views that represent a certain relation between the ontology components or the related ontologies in one table instead of having one table per view.
- Storing the expert's subjective view about each used fuzzy number and linguistic term.
- During insertion time any inconsistent input will be rejected. A clear message showing the reason of rejection is returned to user.

The rest of this section explains this contribution in more details.

Firstly, expert defines his subjective view about all fuzzy numbers using *ExpertFuzzyNumber* table. According to these definitions he/she can store his fuzzy ontology in our schema.

Figure 3 shows the proposed fuzzy ontology database scheme. The list of all ontologies after stemming is stored in the *ontology* table. List of all stemmed classes are stored in the *class* table. *OntologyClasses* table specifies each ontology classes. *FuzzyClassRelation* table stores the fuzzy relation between classes in multi view, e.g., 'sub', 'sup', 'equiv', 'related to', and 'disjoint'. The record {c1, rel, c2, mship_deg1, ..., mship_degn, is_inferred} means that class c1 is related to class c2 with the relation rel with degree mship_deg1 according to view 1 and with degree mship_degn according to view n. An *is_inferred* attribute takes the value 0 or 1 to specify whether or not this record is inferred. *Mship_deg* is the result of defuzzifying the fuzzy number entered by the domain expert. The definition of this fuzzy number is taken from *ExpertFuzzyNumber* table. Then, COA defuzzification method is used to defuzzify this fuzzy number.

Property table stores all properties, i.e., object and datatype. It differentiates between them using the *is_object* attribute. *FuzzyObjectProperty* table stores the domain and the range of each object property (OP) with a fuzzy number expresses to what extent a certain domain has a certain object property with a certain range. *InversreObjectProperty* table specifies which object property is inverse. *FuzzyObjectPropertyRelation* table used to store the fuzzy relation between object properties in multi view, e.g., 'sub', 'sup', 'disjoint', 'related to', 'equiv'. The fuzzy relation entered by the expert is a fuzzy number, which is stored in the table after defuzzifying it.

ObjectPropertyCharacteristics table define each object property (op) characteristics, e.g., functional, transitive. While, op restrictions are defined for each of them and stored in *objectPropertyRestriction* table. *FuzzyDataTypeProperty* table stores classes that have this datatype property and specify the datatype of its value. Datatype property restrictions and characteristics are stored in *dataTypeRestriction* table and *dataTypeCharacteristics* table respectively.

Term table stores all terms. *FuzzyTermRelation* table stores the fuzzy relation between terms in multi view, e.g., 'equiv', 'related to', and 'disjoint'. The record {t1, rel, t2, mship_deg1... mship_degn, is_inferred} means that term t1 is related to term t2 with the relation rel with degree mship_deg1 according to view 1 and degree mship_degn according to view n. The is_inferred attribute takes the value 0 or 1 to specify whether or not this record is inferred. Mship_deg is the result of defuzzifying the fuzzy number entered by the domain expert.

FuzzyClassTerm table stores each class terms. The record (c, t, mship_deg1, ..., mship_degn, is_inferred), means that term t can express class c with degree mship_deg1 according to view 1 and with degree mship_degn according to view n. The is_inferred attribute specifies whether or not this record is inferred.

Individual table stores all individuals. *FuzzyIndividualRelation* table stores the fuzzy relation between individuals in multi view, e.g., 'equiv', 'disjoint', 'related to'. *FuzzyClassIndividual* table stores each class individuals. *FuzzyIndividualObjectProperty* table stores object property that each individual has and its value (another individual).

Table2 shows a comparison between the proposed model and three other models. Compared with the other models, the proposed model assumes fuzzy ontology components as concepts, terms, properties, relations. Relations are represented using fuzzy numbers and linguistic terms, which are defined according to experts' subjective view. Each relation can have more than one value each reflect a certain view. All fuzzy ontology components are stored in a relational database after stemming.

Table2: comparison between the proposed model and other models according to fuzzy ontology components and the used storing technique

Model Name	Fuzzy Ontology Components												Storing Fuzzy Ontology	
	Concept		Property		Term		Individual		Relations				Files	1
									Crisp	Fuzzy Relation				
										Membership degree	Number Fuzzy			
	stem	not	stem	not	stem	not	stem	not				Static Definition	Expert Definition	
the proposed model	✓		✓		✓		✓		✓	✓		✓		
F. Zhang		✓		✓				✓	✓	✓				✓
FORM		✓				✓			✓	✓			✓	
L. Al-Jadir		✓		✓					✓					✓

6. 'Computational Intelligence' ontology: a case study

This section applies the proposed model on computational intelligence ontology stated in FORM as a case study [20]. As shown in figure 4, this ontology has three categories, five keywords. Each keyword represents one or more category through relation with membership degree value. Here, we represent this relation using fuzzy numbers instead of membership value.

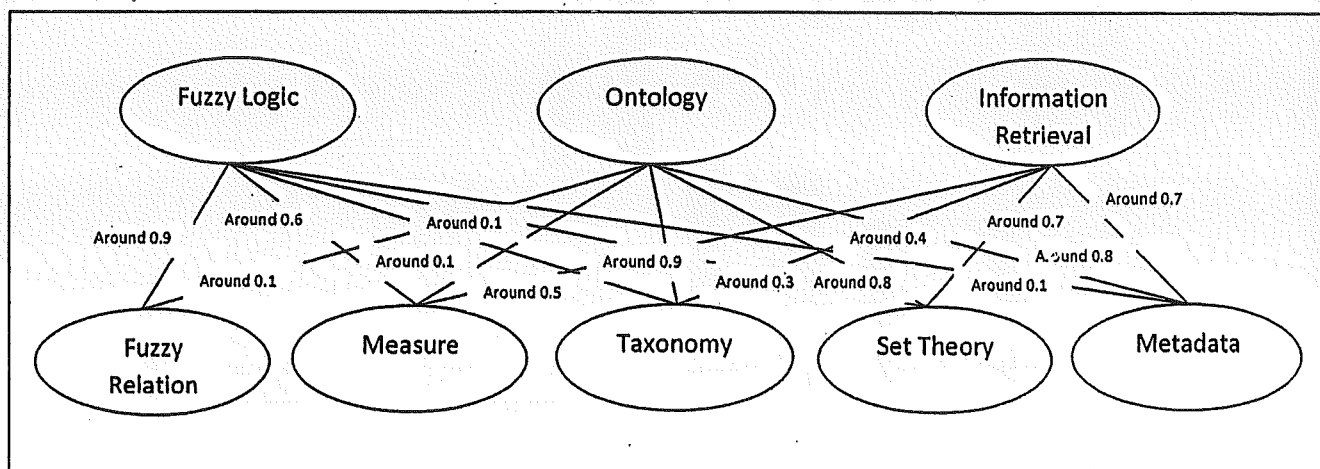


Fig.4: applying the proposed model on 'Computational Intelligence ontology

For storing this ontology according to the proposed model, two steps are required as follows:

1- Firstly, an expert stores his fuzzy number definitions, see figure 5.

2- Lastly, he invokes the following procedure to insert such relations.

```
insert_fuzzy_class_term ('Fuzzy Relation', 'Fuzzy Logic', 'around', 0.9, 'theoretical point of view');
insert_fuzzy_class_term ('Fuzzy Relation', 'ontology', 'around', 0.1, 'theoretical point of view');
insert_fuzzy_class_term ('Measure', 'Fuzzy Logic', 'around', 0.6, 'theoretical point of view');
insert_fuzzy_class_term ('Measure', 'Information Retrieval', 'around', 0.5, 'theoretical point of view');
```

EXPERTFUZZYNUMBER			
EDIT	FUZZY NUM ID	FUZZY NUM NAME	MEMBERSHIP FUNCTION
	3	not less than	trapizoidal
	1	around	triangular
	2	about	pl
row(s) 1 - 3 of 3			

TRIANGULARFUZZYNUMBER			
EDIT	N ALPHA	P ALPHA	FUZZY NUM ID
	.1	.1	1

Fig.5: fuzzy numbers defined by expert

This procedure will insert each tuple after interpreting the entered fuzzy number according to the expert definition then defuzzify it using COA defuzzification method. After that, it checks the tuple consistency. Each consistent tuple is inserted and other ones are rejected, showing the expert the rejection reason. The data inserted in database should be like figure 6.

FUZZYCLASSTERM					
ID	CLASS	TERM	MEMBERSHIP	IS_INFERRED	VIEW ID
	4	12	.4	0	3
	5	12	.8	0	3
	2	12	.1	0	3
	2	11	.6	0	3
	2	13	.5	0	3
	5	13	.7	0	3
	4	13	.7	0	3
	5	11	.1	0	3
	4	11	.8	0	3
	1	11	.9	0	3
	1	12	.1	0	3
row(s) 1 - 11 of 11					

Fig.6: fuzzyClassTerm table

7. Conclusion & Future Work

In this paper, an enhanced model for storing fuzzy ontology is presented. The proposed model is more flexible and generic to model varieties of ontologies. This model extends previous ones to incorporate concepts, properties, terms, relations and individuals. It also allows experts to define relations as fuzzy numbers instead of a membership values. Also, it helps experts to build their fuzzy related fuzzy ontology according to multi-view in an easy and efficient way, and automatically storing their ontology in relational database. The future direction to work in this area would be: to implement a reasoner that checks the ontology consistency and infers new knowledge, also to build a query answering tool.

8. References

- [1] C. D. Barranco, J. R. Campana, J. M. Medina, O. Pons, "On Storing Ontologies including Fuzzy Datatypes in Relational Databases," IEEE, 2007.
- [2] F. Bobillo, M. Delgado, J. Gómez-Romero, "DeLorean: A reasoner for fuzzy OWL 2," Expert Systems with Applications: An International Journal Volume 39 Issue 1, pp. 258-272, January, 2012
- [3] I. Astrova, N. Korda, A. Kalja, "Storing OWL Ontologies in SQL Relational Databases," World Academy of Science, Engineering and Technology 29, 2007
- [4] J. Zhai, Y. Liang, Y. Yu, J. Jiang, "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce," JOURNAL OF SOFTWARE, VOL. 3, NO. 9, DECEMBER 2008.
- [5] 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.
- [6] Y. Lv, Z. M. Ma, N. Zhang, "Fuzzy ontology storage in fuzzy relational DB," IEEE, sixth international conference on fuzzy system and knowledge discovery, 2009.
- [7] F. Zhang, Z.M. Ma, L. Yan, J. Cheng, "Storing Fuzzy Ontology in Fuzzy Relational Database," DEXA 2011, Part II, LNCS 6861, pp. 447-455, springer, 2011.
- [8] F. Zhang, Li Yan, Z. M. Ma, "Reasoning of fuzzy relational databases with fuzzy ontologies," International Journal of Intelligent Systems, Volume 27, Issue 6, pages 613-634, June 2012.
- [9] L. Al-Jadir, C. Parent, S. Spaccapietra, "Reasoning with large ontologies stored in relational databases: The OntoMinD approach," Data & Knowledge Engineering 69, pp. 1158-1180, 2010.
- [10] F. B. Ortenga, M. Delgado Calvo-Flores, "Managing Vagueness in Ontologies," PHD Dissertation, Granada, October 2008
- [11] 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.
- [12] T. T. Quan, S. C. Hui, A. C. M. Fong, T. H. Cao, "Automatic Fuzzy Ontology Generation for Semantic Web," IEEE Transaction on Knowledge and Data Engineering 18.6, pp. 842-856, 2006.
- [13] 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.
- [14] H. M. Gu, X. Wang, Y. Ling, J. Q. Shi, "Building a Fuzzy Ontology of Edutainment Using OWL," the 7th International Conference on Computational Science (ICCS 2007), pp. 591-594, 2007.

- [15] 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.
- [16] C. M. A. Yeung, H. F. Leung, "Ontology with Likeliness and typicality of Objects in Concepts," the 25th International Conference on Conceptual Modelling (ER 2006), Vol. 4215, pp. 98-111, 2006.
- [17] M. Abulaish, L. Dey, "Interoperability among Distributed Overlapping Ontologies- A Fuzzy Ontology framework," IEEE/WIC/ ACM International Conference on Web Intelligence (WI 2006), pp. 397-403, 2006.
- [18] M. Abulaish, L. Dey, "A Fuzzy Ontology Generation Framework for Handling Uncertainties and Nonuniformity in Domain Knowledge Description," International Conference on computing: Theory and Application(ICCTA 2007), pp.287-293, 2007.
- [19] M. A. A. Leite and I. L. M. Ricarte," A Framework for Information Retrieval Based on Fuzzy Relations and Multiple Ontologies," Springer, pp. 292-301,2008.
- [20] R. Pereira, I. Ricarte, F. Gomide, " Information Retrieval with FROM: The Fuzzy Relational Ontological Model," INTERNATIONAL JOURNAL OF INTELLEAGENT SYSTEMS, VOL. 24, 340-356, 2009.
- [21]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.

Granular Computing Approach for Manipulating Fuzziness Uncertainty

Assem Ahmed Alsawy⁽¹⁾

Hesham Ahmed Hefny⁽²⁾

Abstract

Basic ideas of crisp information granulation have appeared in various fields, such as interval analysis, data quantization, cluster analysis, machine learning, databases, and many others. The topic of fuzzy information granulation was first proposed and discussed by Zadeh in 1979. There is a fast growing and renewed interest in this topic. Granular computing is likely to play an important role in the evolution of fuzzy logic and its applications. This paper explores a view of granular computing as a human-inspired computational paradigm to introduce a new representation for fuzzy numbers and shows the benefit of adopting a unified granular form of representation.

Keywords—Granularity, Granular Computing, Unified Granular Number, Fuzzy logic, Fuzzy Number

1. Introduction

The information that people obtain is usually uncertain and incomplete. In recent years, the definition of inexact number has been emerged from different perspectives. Young R.C. (1931) and Moore R E. (1979) proposed the theory of interval numbers [1-2]. Zadeh (1965) proposed fuzzy set theory [3], then he introduced type-2 fuzzy sets , and even higher types, in 1975 [4]. Pawlak (1982) proposed the rough sets theory [5]. Deng. (1982) proposed the grey systems theory [6]. Atanassov (1983) suggested the theory of Intuitionistic fuzzy sets [7]. Buehrer and. Gau (1992) proposed the theory of vague sets [8]. All the previously proposed sets theories are constructs falling under the same umbrella of Granular Computing for handling various types of uncertainty. The theory of granular computing (GrC) has recently emerged as a coherent conceptual and algorithmic platform aimed at the representation and processing of information granules [9].

⁽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

At present, granular computing is more a theoretical perspective than a well established set of methods or principles [10]. As a theoretical perspective, it encourages an approach that recognizes and exploits the presented knowledge in data at various levels of resolution or scales. In this sense, it encompasses all methods which provide flexibility and adaptability in the resolution at which knowledge or information is extracted and represented [11].

2. Related work

There are many forms of imprecise granular numbers such as: interval numbers, fuzzy numbers, rough numbers, vague numbers, gray numbersetc. Classical approach for information processing requires computers (as being algorithmic machines) to use exact and precise algorithms that manipulate only confident data and precise numbers. There is an urgent need for new generations of algorithms which can handle real world problems containing various forms of uncertain data and imprecise numbers of different types. This is a challenging work that needs much of research efforts. Some researchers tried to create formulas of transformations among different representations of uncertain granular numbers [12]. Figure 1 shows what has been accomplished so far in that approach.

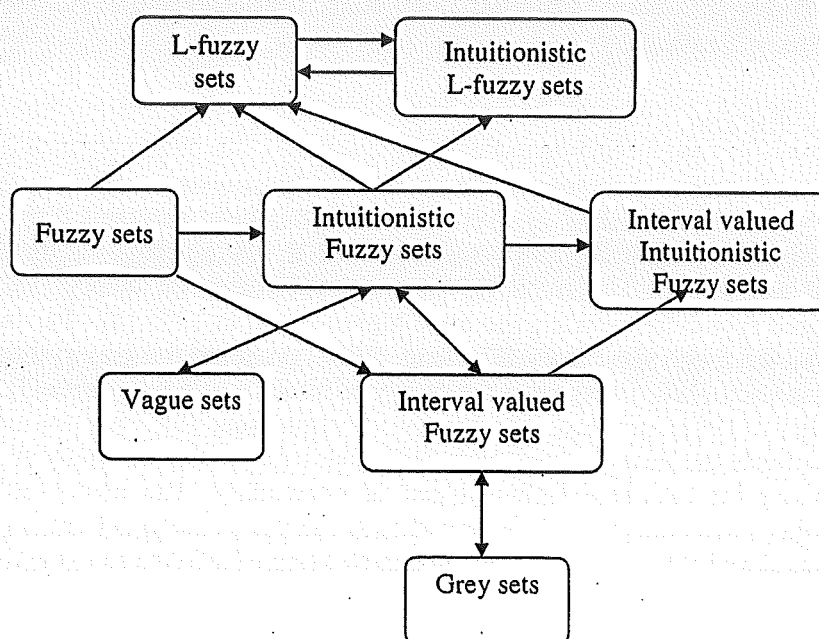


Figure 1: Links between some theories modeling imprecision and uncertainty [12].

It should be noted that, even all transformational formulas among all types of representation of uncertain numbers, it will be so difficult to various kinds of problems containing different types of uncertain granular numbers.

3. Granular Computing

Granular computing (GrC) is an emerging computing paradigm of information processing. It concerns the processing of complex information entities called information granules, which arise in the process of data abstraction and derivation of knowledge from information [15]. Generally speaking, information granules are collections of entities that usually originate at the numeric level and are arranged together due to their similarity, functional or physical adjacency, indistinguishability, coherency, or the like [16].

The basic concepts of granulation form the core of fuzzy logic; these concepts are the principal distinguishing features of fuzziness. More specifically, in fuzzy logic everything is or is allowed to be granulated, the idea of fuzzy information granulation is inspired by the ways in which humans granulate information and reason with it [14].

Granular computing has begun to play important role in bioinformatics, e-business, security, machine learning, data mining, high-performance computing, optimization and wireless mobile computing in terms of efficiency, effectiveness, robustness and uncertainty [13].

4. Unified Granular Number

In this work we suggest a Unified Granular Number (UGN), that we call it G-Number to be as general form for any uncertain granular number. G-Number represents a higher level of abstract that hold common properties of all uncertain granular numbers while ignoring some particular properties which are not necessary to be considered in such higher abstract level. The main benefit of using such a proposed G- number is the ability to represent all types of granular numbers using unified formality that greatly simplifies arithmetic operations.

4.1 Definition: G- Number is an extension of a regular number in the sense that it does not refer to one single value but rather it refers to a connected set of possible values. G- Number is a quantity whose value is imprecise, rather than exact or accurate as the case with "ordinary" (single-valued) numbers.

G-number X can be expressed by two factors which are: the center value of number x , and radius of number r_x as in form $X = G(x, r_x)$

This formulation of G-number represents the most common properties of uncertain granular number which are lower and upper boundaries. In this paper, we show the process of transforming various shapes of type-1 fuzzy numbers to the proposed form of G-Number.

4.2 Triangular Fuzzy Number

For a triangular fuzzy number X , as shown in figure 2, which can be denoted by

$$X = Tri(a, b, c)$$

With membership function $\mu_X(w) = \max\left(\min\left(\frac{x-a}{b-a}, \frac{c-x}{c-b}\right), 0\right)$

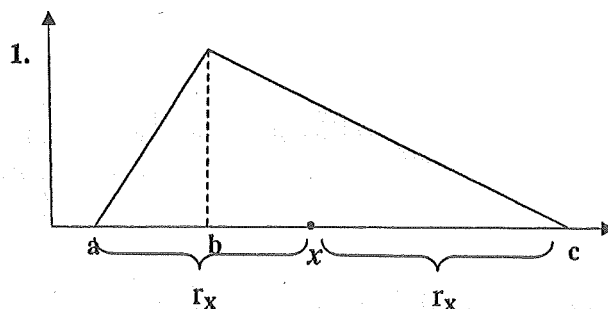


Figure 2: Triangular fuzzy number

Since x is midpoint of $\overline{ac}$, then according to the newly proposed notation we have:

$$X = G(x, r_x) = \{w, x - r_x \leq w \leq x + r_x, 0 \leq \mu_X(w) \leq 1\}$$

$$a = x - r_x, c = x + r_x$$

$$\text{Also } x = \frac{a+c}{2} \text{ and } r_x = \frac{c-a}{2} \quad \text{for all } a, b, c, x, r_x \geq 0$$

$$Tri(a, b, c) = G(x, r_x)$$

4.3 Trapezoidal Fuzzy Number

For a trapezoidal fuzzy number Y , as shown in figure 3, which can be denoted by:

$$Y = Trap(a, b, c, d)$$

With membership function $\mu_Y(w) = \max\left(\min\left(\frac{x-a}{b-a}, 1, \frac{c-x}{c-b}\right), 0\right)$

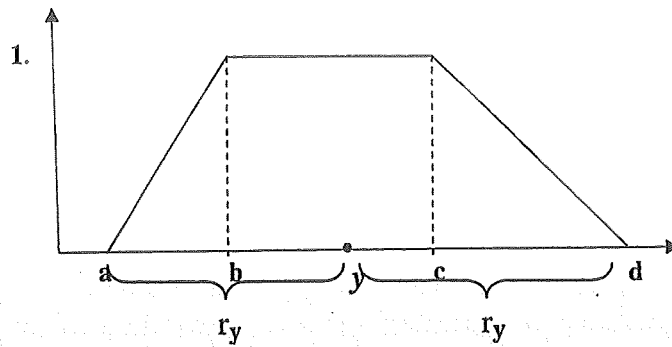


Figure 3: Trapezoidal fuzzy number

Since y is midpoint of $\overline{ad}$, then according to the newly proposed notation we have:

$$Y = G(y, r_y) = \{w, y - r_y \leq w \leq y + r_y, 0 \leq \mu_Y(w) \leq 1\}$$

$$a = x - r_y, \quad d = x + r_y,$$

$$\text{Also } y = \frac{a+d}{2} \quad \text{and} \quad r_y = \frac{d-a}{2} \quad \text{for all } a, b, c, d, x, r_y \geq 0$$

$$Y = \text{Trap}(a, b, c, d) = G(y, r_y)$$

4.4 Gaussian Fuzzy Number

For Gaussian fuzzy number Z , as shown in figure 4, which can be denoted by:

$$Z = \text{Gaussian}(\mu, \sigma)$$

With membership function $\mu_Z(w) = e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$

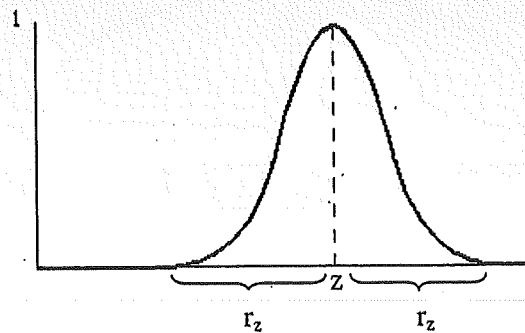


Figure 4: Gaussian fuzzy number

$$Z = G(z, r_z) = \{w, z - r_z \leq w \leq z + r_z, 0 \leq \mu_z(w) \leq 1\}$$

$$z = \mu \quad \text{and} \quad r_z = 3\sigma$$

$$Z = \text{Gaussian}(\mu, \sigma) = G(z, r_z)$$

4.5 Ellipse –Shaped Fuzzy Number

Polygon fuzzy number is a general name for any fuzzy number consisting of a number of points (called vertices) and number of line segments. Triangular fuzzy number and trapezoidal fuzzy number are two special cases of polygon fuzzy number [17]. For a huge number of vertices, the polygon fuzzy number turned to be can be approximated to a continuous convex shape, e.g. circular or ellipse –shaped fuzzy number. Of course, ellipse –shaped fuzzy number is also considered a general form for circular fuzzy number.

For ellipse fuzzy number S , as shown in figure 5, which can be denoted by:

$$S = \text{ellipse}(a, b, x_0, y_0).$$

With membership function $\mu_S(w) = y_0 + b \sqrt{1 - \left(\frac{x-x_0}{a}\right)^2}$

In case of $y_0 = 0$ and $b = 1$. Then, it is enough to form the number S as follows:

$$S = \text{ellipse}(x, a)$$

With membership function $\mu_S(w) = \sqrt{1 - \left(\frac{x-x_0}{a}\right)^2}$

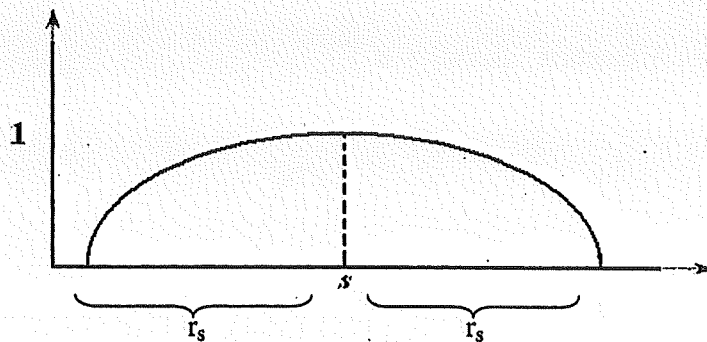


Figure 5: Ellipse –Shaped Fuzzy number

As shown in figure 5, the Ellipse –Shaped Fuzzy number can be represented using the notation of G-number as follows:

$$S = G(s, r_s) = \{w, s - r_s \leq w \leq s + r_s, 0 \leq \mu_s(w) \leq 1\}$$

with $s = x$ and $r_x = a$

$$S = ellipse(x, a) = G(s, r_s)$$

One may think that, the weakness point of using the above formula of unified granular number to represent various types of fuzzy numbers is the neglecting of the membership function $\mu(w)$ which is the main idea in fuzzy set, the answer is: according to philosophy of defuzzification, a single crisp value that depends on the shape of the membership function, is extracted from the support of the fuzzy set to act as a unique representation to the whole fuzzy set (i.e. fuzzy number) from a decision point of view. Following the same philosophy, the proposed notation of G-Number represents the fuzzy number by using two parameters that represent the center and radius of the support (granular) respectively [18].

However, a modified formula of the G- Number may be more appropriate by adding a new factor that represent the weight uncertainty in the granule, which is fairly represented by its area.

Thus, the modified proposed notation of the G-Number includes the weight of the number X which is its area $A(X)$. Thus we have the final formula to be adopted of G-Number as:

$$X = G \frac{(x, r_x)}{A(X)},$$

The following cases are consequently arise:

- For triangular fuzzy number $X = Tri(a, b, c)$, see figure 2, the new formula for equivalent unified granular number is:

$$X = G \frac{(x, r_x)}{A(X)} \quad A(X) = r_x$$

- For trapezoidal fuzzy number $Y = Trap(a, b, c, d)$, see figure 3, the new formula for equivalent unified granular number is:

$$Y = G \frac{(y, r_y)}{A(Y)} \quad A(Y) = r_y + \frac{1}{2}(c - b)$$

- For Gaussian fuzzy number $Z = Gaussian(\mu, \sigma)$, see figure 4, the new formula for equivalent unified granular number is:

$$Z = G \frac{(z, r_z)}{A(Z)} \quad A(Z) = \sigma \sqrt{\pi}$$

- For Ellipse -Shaped Fuzzy number $S = ellipse(x, a)$, see figure 5, the new formula for equivalent unified granular number is:

$$S = G \frac{(s, r_s)}{A(S)} \quad A(S) = \frac{1}{2} \pi a$$

The main idea behind creating G-numbers is to have a general form of all types of uncertain granular numbers. Having a unified form for representing uncertain numbers allows performing arithmetic operations in a simple way regardless of the shape of uncertain numbers. In the following sections we introduce some proved arithmetic operations that we can use in calculation on G-numbers.

5. Operation on G-number

The most required operations so far are addition and subtraction

5.1 Addition of G-numbers

For any two G- numbers X and Y the addition operation is defined as follows:

$$G \frac{(x, r_x)}{A(X)} + G \frac{(y, r_y)}{A(Y)} = G \frac{(x+y, r_x+r_y)}{A(X)+A(Y)}$$

To prove this formula, we check it for each type of uncertain granular number:

1. In case of triangular fuzzy numbers

$$G \frac{(x, r_x)}{A(X)} = \text{Tri}(x - r_x, b_1, x + r_x) \text{ and } A(X) = r_x \text{ as in figure 6}$$

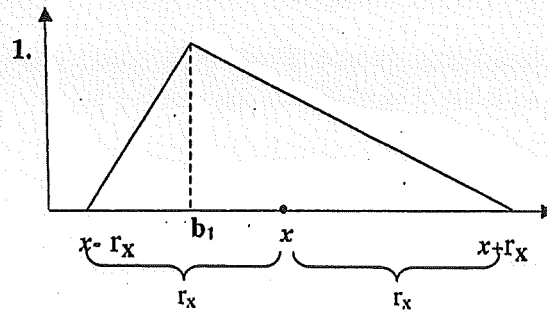


Figure 6: $X = \text{Tri}(x - r_x, b_1, x + r_x)$

$$G \frac{(y, r_y)}{A(Y)} = \text{Tri}(y - r_y, b_2, y + r_y) \text{ and } A(Y) = r_y \text{ as in figure 7}$$

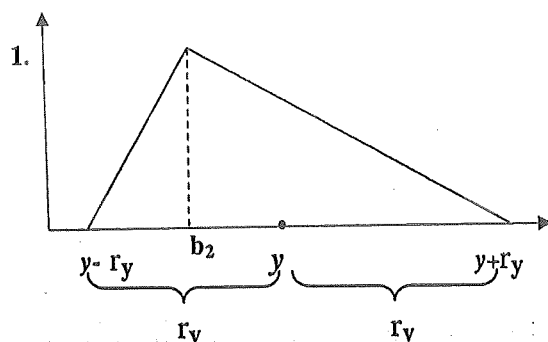


Figure 7: $Y = \text{Tri}(y-r_y, b_2, y+r_y)$

$$\begin{aligned}
 G \frac{(x, r_x)}{A(X)} + G \frac{(y, r_y)}{A(Y)} &= \text{Tri}(x-r_x, b_1, x+r_x) + \text{Tri}(y-r_y, b_2, y+r_y) \\
 &= \text{Tri}((x-r_x)+(y-r_y), b_1+b_2, (x+r_x)+(y+r_y)) \\
 &= \text{Tri}((x+y)-(r_x+r_y), b_1+b_2, (x+y)+(r_x+r_y)) \quad , \text{ see figure 8} \\
 &= G \frac{(x+y, r_x+r_y)}{A(X+Y)}
 \end{aligned}$$

Where,

$$A(X+Y) = r_x + r_y = A(X) + A(Y)$$

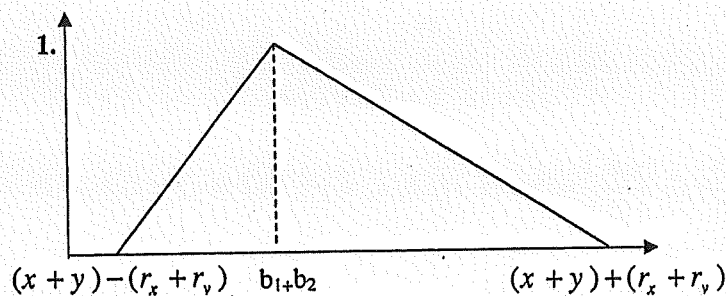


Figure 8: $X + Y$

2. In case of trapezoidal fuzzy numbers

$$G \frac{(x, r_x)}{A(X)} = \text{Trap}(x-r_x, b_1, c_1, x+r_x) \quad \text{and} \quad A(X) = r_x + 1/2(c_1 - b_1) \quad \text{as in figure 9}$$

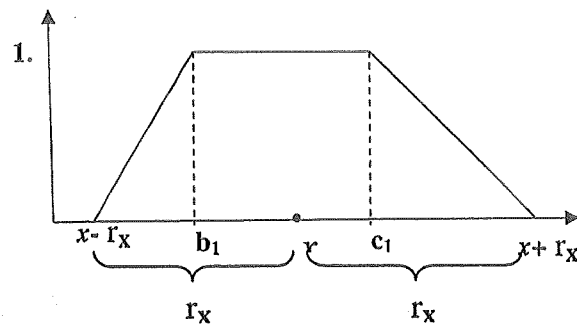


Figure 9: $X = \text{Trap}(x-r_x, b_1, c_1, x+r_x)$

$$G \frac{(y, r_y)}{A(Y)} = \text{Trap}(y-r_y, b_2, c_2, y+r_y) \quad \text{and} \quad A(Y) = r_y + 1/2(c_2 - b_2) \quad \text{as in figure 10}$$

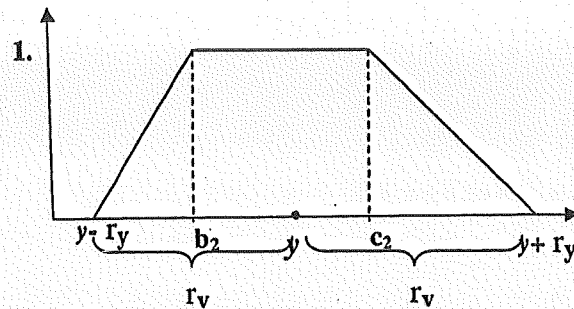


Figure 10: $Y = \text{Trap}(y-r_y, b_2, c_2, y+r_y)$

$$\begin{aligned} G \frac{(x, r_x)}{A(X)} + G \frac{(y, r_y)}{A(Y)} &= \text{Trap}(x-r_x, b_1, c_1, x+r_x) + \text{Trap}(y-r_y, b_2, c_2, y+r_y) \\ &= \text{Trap}((x-r_x) + (y-r_y), b_1+b_2, c_1+c_2, (x+r_x) + (y+r_y)) \\ &= \text{Trap}((x+y) - (r_x+r_y), b_1+b_2, c_1+c_2, (x+y) + (r_x+r_y)) \quad \text{as shown in figure 11} \\ &= G \frac{(x+y, r_x+r_y)}{A(X+Y)} \end{aligned}$$

Where,

$$\begin{aligned} A(X+Y) &= r_x + r_y + 1/2((c_1+c_2) - (b_1+b_2)) \\ &= r_x + 1/2(c_1-b_1) + r_y + 1/2(c_2-b_2) \\ &= A(X) + A(Y) \end{aligned}$$

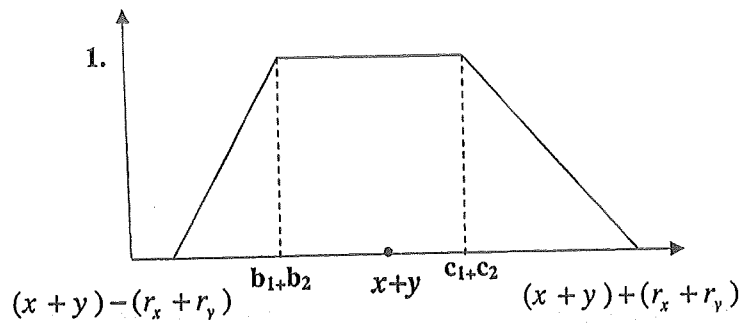


Figure 11: X + Y

5.2 Subtraction of G-numbers

For any two G-Number X and Y the subtraction operation is defined as Follows:

$$G \frac{(x, r_x)}{A(X)} - G \frac{(y, r_y)}{A(Y)} = G \frac{(x-y, r_x+r_y)}{A(X)+A(Y)}$$

To prove this formula, we check it for each type of uncertain granular number:

1. In case of triangular fuzzy number

$$G \frac{(x, r_x)}{A(X)} = Tri(x - r_x, b_1, x + r_x) \text{ and } A(X) = r_x \text{ as in figure 6}$$

$$G \frac{(y, r_y)}{A(Y)} = Tri(y - r_y, b_2, y + r_y) \text{ and } A(Y) = r_y \text{ as in figure 7}$$

$$\begin{aligned} G \frac{(x, r_x)}{A(X)} - G \frac{(y, r_y)}{A(Y)} &= Tri(x - r_x, b_1, x + r_x) - Tri(y - r_y, b_2, y + r_y) \\ &= Tri((x - r_x) - (y + r_y), b_1 - b_2, (x + r_x) - (y - r_y)) \\ &= Tri((x - y) - (r_x + r_y), b_1 - b_2, (x - y) + (r_x + r_y)) \text{ as in figure 12} \\ &= G \frac{(x - y, r_x + r_y)}{A(X - Y)} \end{aligned}$$

where,

$$A(X - Y) = r_x + r_y = A(X) + A(Y)$$

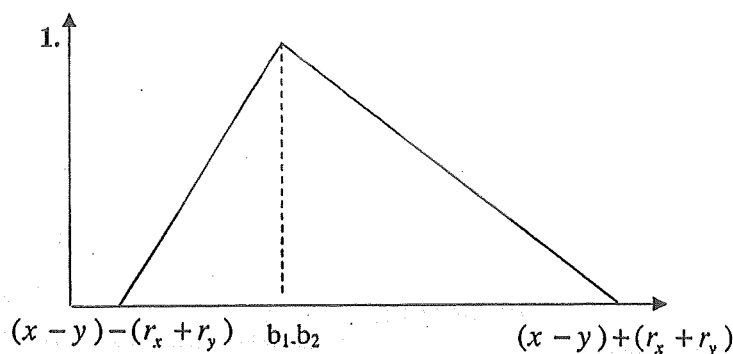


Figure 12: X - Y

2. In case of trapezoidal fuzzy number

$$G \frac{(x, r_x)}{A(X)} = \text{Trap}(x - r_x, b_1, c_1, x + r_x) \text{ and } A(X) = r_x + 1/2(c_1 - b_1) , \text{ see figure 9}$$

$$G \frac{(y, r_y)}{A(Y)} = \text{Trap}(y - r_y, b_2, c_2, y + r_y) \text{ and } A(Y) = r_y + 1/2(c_2 - b_2) , \text{ see figure 10}$$

$$\begin{aligned} G \frac{(x, r_x)}{A(X)} - G \frac{(y, r_y)}{A(Y)} &= \text{Trap}(x - r_x, b_1, c_1, x + r_x) - \text{Trap}(y - r_y, b_2, c_2, y + r_y) \\ &= \text{Trap}((x - r_x) - (y + r_y), b_1 - c_2, c_1 - b_2, (x + r_x) - (y - r_y)) \\ &= \text{Trap}((x - y) - (r_x + r_y), b_1 - c_2, c_1 - b_2, (x - y) + (r_x + r_y)) , \text{ see figure 13} \\ &= G \frac{(x - y, r_x + r_y)}{A(X - Y)} \end{aligned}$$

where,

$$\begin{aligned} A(X - Y) &= r_x + r_y + 1/2((c_1 + c_2) - (b_1 + b_2)) \\ &= r_x + 1/2(c_1 - b_1) + r_y + 1/2(c_2 - b_2) = A(X) + A(Y) \end{aligned}$$

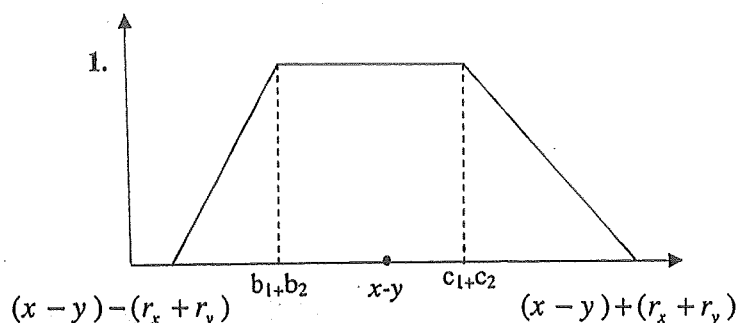


Figure 13: X - Y

5.3 G-Number as a Triangular Fuzzy number

To transform a unified granular G-number back to a triangular fuzzy number, i.e.

$$X = G \frac{(x, r_x)}{A(X)} = Tri(a, b, c)$$

Then, $a = x - r_x$, $c = x + r_x$, and $x - r_x \leq b \leq x + r_x$.

It is clear that infinite numbers of different triangles may be obtained according to different values of b . Although we can not estimate exactly the original value of b , which is the cost paid for a unification granular form, yet accepted estimate values for b can be obtained in terms of the centroid of the triangle, as shown in figure 14. The centroid of any triangle fuzzy number $Tri(a, b, c)$ is given by the following formula [19]:

$$Centroid(X) = \frac{a+b+c}{3} = \frac{x-r_x+b+x+r_x}{3} = \frac{2x+b}{3}$$

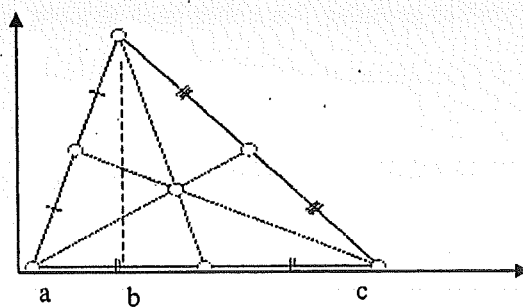


Figure 14: Centroid of Triangular Fuzzy Number

Therefore $x - \frac{1}{3}r_x \leq \text{Centroid}(X) \leq x + \frac{1}{3}r_x$ as shown in Figure 15

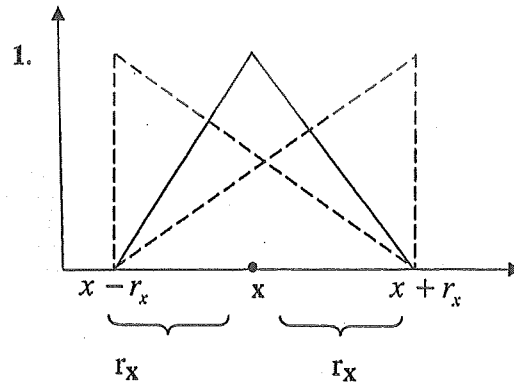


Figure 15: Range of Triangular Fuzzy Number

Therefore, the accepted choices for b values lie within the following interval:

$$b \in \left[x - \frac{1}{3}r_x, x + \frac{1}{3}r_x \right]$$

In case of balanced triangular fuzzy number

$$a = x - r_x, b = x, \text{ and } c = x + r_x$$

$$\text{Centroid}(X) = \frac{a+b+c}{3} = \frac{x - r_x + x + x + r_x}{3} = x$$

5.4 G-Number as a Trapezoidal Fuzzy Number

To transform a unified granular G-number back to a trapezoidal fuzzy number, i.e.

$$X = G \frac{(x, r_x)}{A(X)} = \text{Trap}(a, b, c, d)$$

$$a = x - r_x, d = x + r_x,$$

if $A(X) = r_x$, that means $c = b$ so the trapezoid turns to triangle

if $A(X) > r_x$ then $x - r_x \leq b < c \leq x + r_x$

$$\text{Since } A(X) = \frac{1}{2}(2r_x + (c - b))$$

$$\text{So } (c - b) = 2(A(X) - r_x)$$

$$\text{Then } x - r_x \leq b < b + 2[A(x) - r_x] \leq x + r_x$$

$$\text{and } x - r_x \leq c - 2[A(x) - r_x] < c \leq x + r_x$$

Again, among infinite possible choices of trapezoidal fuzzy numbers for the different values of b and c , an accepted choice values for b and c can be obtained in terms of the centroid of the trapezoid.

The general formula of the centroid of a trapezoid fuzzy number $Trap(a, b, c, d)$ is [19]:

$$\begin{aligned} Centroid(X) &= \frac{1}{3} \left[a + b + c + d - \frac{dc - ab}{(d + c) - (a + b)} \right] \\ &= \frac{1}{3} \left[2x + (c + b) \frac{(c - b) + r_x}{(c - b) + 2r_x} \right] \\ &= \frac{1}{3} \left[2x + (c + b) \frac{2A(X) - r_x}{2A(X)} \right] \\ &= \frac{1}{3} \left[2x + (c + b) \frac{2A(X) - r_x}{2A(X)} \right] \\ &= \frac{1}{3} \left[2x + (c + b) \left(1 - \frac{r_x}{2A(X)} \right) \right] \\ &= \frac{1}{3} [2x + K] \end{aligned}$$

Accepted choices for K values lie within the following interval:

$$K \in \left[x - \frac{1}{3}r_x, x + \frac{1}{3}r_x \right]$$

Thus,

$$(c + b) \left(1 - \frac{r_x}{2A(X)} \right) \in \left[x - \frac{1}{3}r_x, x + \frac{1}{3}r_x \right]$$

$$2(b + A(X) - r_x) \left(1 - \frac{r_x}{2A(X)} \right) \in \left[x - \frac{1}{3}r_x, x + \frac{1}{3}r_x \right]$$

$$(b + A(X) - r_x) \in \left(\frac{A(X)}{2A(X) - r_x} \right) \left[x - \frac{1}{3}r_x, x + \frac{1}{3}r_x \right]$$

$$b \in \left(\frac{A(X)}{2A(X) - r_x} \right) \left[x - \frac{1}{3}r_x, x + \frac{1}{3}r_x \right] - (A(X) + r_x)$$

Then,

$$C \in \left(\frac{A(X)}{2A(X) - r_x} \right) \left[x - \frac{1}{3}r_x, x + \frac{1}{3}r_x \right] + (A(X) - 3r_x)$$

In case of balanced trapezoidal fuzzy number as shown in figure 16

$$a = x - r_x, d = x + r_x$$

$$\text{since } A(X) = r_x + r'_x$$

$$b = x - r'_x = x + r_x - A(X), \text{ and } c = x + r'_x = x - r_x + A(X),$$

$$\text{Centroid}(X) = x$$

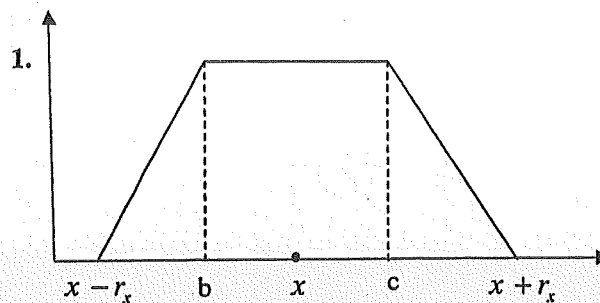


Figure 16: balanced Trapezoidal Fuzzy Number

6. Illustrative Example

This example shows how G-Numbers could replace different types of fuzzy numbers. The example is a 4-city traveling salesman problem. Let the salesman has to visit all cities and return to home at the end of journey. The set of cities is $C = \{C_1, C_2, C_3, C_4\}$ as shown in figure 17

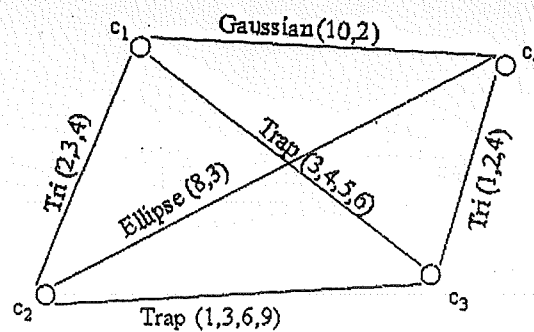


Figure 17: Distance between cities as fuzzy numbers

D : is the distance function and it can be defined as following:

$D : C \times C \rightarrow F$ Since F is a set of all fuzzy numbers

$$D(C_1, C_2) = \text{Tri}(2, 3, 4)$$

$$D(C_1, C_3) = \text{Trap}(3, 4, 5, 6)$$

$$D(C_1, C_4) = \text{Gaussian}(10, 2)$$

$$D(C_2, C_3) = \text{Trap}(1, 3, 6, 9)$$

$$D(C_2, C_4) = \text{ellipse}(8, 3)$$

$$D(C_3, C_4) = \text{Tri}(1, 2, 4)$$

To find the shortest cycle, we may use the representation of granular number as follows:

$$\text{Tri}(2, 3, 4) = G \frac{(3, 1)}{1}$$

$$\text{Trap}(3, 4, 5, 6) = G \frac{(4.5, 1.5)}{2}$$

$$\text{Gaussian}(10, 2) = G \frac{(10, 6)}{2\sqrt{\pi}}$$

$$\text{Trap}(1, 3, 6, 9) = G \frac{(5, 4)}{5.5}$$

$$\text{ellipse}(8, 3) = G \frac{(8, 3)}{1.5\pi}$$

$$\text{Tri}(1, 2, 4) = G \frac{(2.5, 1.5)}{1.5}$$

By calculating the lengths of all possible paths then ranking them according to the value center point, we can say the shortest path is:

$$C_1 \rightarrow C_2 \rightarrow C_4 \rightarrow C_3 \rightarrow C_1$$

The length of this path can be calculated as following:

$$G \frac{(3, 1)}{1} + G \frac{(8, 3)}{1.5\pi} + G \frac{(2.5, 1.5)}{1.5} + G \frac{(4.5, 1.5)}{2} = G \frac{(18, 7)}{9.212}$$

This number can be converted to any of fuzzy numbers and determine its centroid here we convert it to a balanced trapezoidal fuzzy number

$$G \frac{(18, 7)}{9.212} = \text{Trap}(a, b, c, d)$$

$$a = 18 - 7 = 11$$

$$b = 18 + 7 - 9.212 = 15.788$$

$$c = 18 - 7 + 9.212 = 20.212$$

$$d = 18 + 7 = 25$$

$$\text{Therefore } G \frac{(18, 7)}{9.212} = \text{Trap}(11, 15.788, 20.212, 25)$$

The single value which reflects the total distance is: 18 which is the value in case of crisp numbers.

7. Conclusion

Unified granular number or G-number is a new representation for different types of fuzzy numbers. The main benefit from using G-number is that it provides the ability to homogenate the manipulation of heterogeneous types of fuzzy numbers via converting each of them into the new representation of G-number, then the same set of arithmetic calculations rules can be utilized.

This research defines basic addition and subtraction operations on G-numbers and presents mathematical proofs in cases of triangular and trapezoidal of fuzzy numbers. However, for other types of fuzzy numbers, it is suitable to prove the formulas of addition and subtraction by using alpha-cut method or approximate each of them to trapezoidal fuzzy number as a general case.

Reference s

- [1] Young, R.C., "The algebra of many-valued quantities", *Annals of Mathematics*, 31, pp. 260–290 (1931).
- [2] Moore, R.E., "Method and Application of Interval Analysis", Prentice Hall, London (1979).
- [3] Zadeh, L.A., "Fuzzy sets", *Information and Control* 8, pp. 338–353 (1965).
- [4] Zadeh, L.A., "The concept of a linguistic variable and its application to approximate reasoning", *Information Sciences*, 8 pp 199–249 (1975).
- [5] Pawlak, Z., "Rough sets", *International Journal of Information & Computer Sciences* 11, pp. 341–356 (1982).
- [6] Deng, J.L., "The Control problem of grey systems", *System & Control Letters*, 1(5), pp. 288–294 (1982).
- [7] Atanassov, K. T., "Intuitionistic fuzzy sets", VII ITKR's Session, pp. 391-396 (1983).
- [8] Buehrer, D. J. "Vague logic: A first-order extension of interval probability theory" in *Proc. IEEE Workshop Imprecise and Approximate Computation*, Phoenix, AZ, pp. 83-87 (1992).
- [9] Pedrycz, W., "Knowledge-based clustering: from data to information granules", J. Wiley, Hoboken, NJ, (2005).
- [10] Pedrycz, W. and Gomide, F., "Fuzzy systems engineering: toward human-centric computing", John Wiley, Hoboken, NJ, (2007).
- [11] Pedrycz, W. and Valente de Oliveira, J., "A development of fuzzy encoding and decoding through fuzzy clustering", *IEEE Transactions on Instrumentation and Measurement*, 57, N.4, pp.829-837, (2008).

- [12] Deschrijver, G. and Kerre, E.E., "On the relationship between some extensions of fuzzy set theory", *Fuzzy Sets and Systems*, 133(2), pp. 227-235 (2003).
- [13] Pedrycz, W., Loia, V. and Senatore, S. " Fuzzy clustering with viewpoints", *IEEE Trans. on Fuzzy Systems*, 161(1) pp. 56-74 (2010).
- [14] Bargiela, A., Pedrycz, W., " Granular Computing: An Introduction", Kluwer Academic Publishers, Boston, (2002).
- [15] Pedrycz, W. and Liu, X., " Axiomatic fuzzy set theory and its applications", Springer Verlag, Berlin Vol. 244, pp. 2475-2486 (2009).
- [16] Zadeh, L.A., "Toward a generalized theory of uncertainty (GTU)", *Information Sciences*, 172, pp. 1-40 (2005).
- [17] Kreith, Frank; Goswami, D., "Circles and Ellipses", *The CRC Handbook Of Mechanical Engineering*. CRC Press. pp. 11-18. (2005).
- [18] Assem A Alsawy and Hesham A Hefny, "On Uncertain Granular Numbers" , *International Journal of Computer Applications* 62(18), pp 20-27,. Foundation of Computer Science, New York, USA, (2013).
- [19] Wang Y. M., J. B. Lang, D. L. Xu and K. S. Chin, "On the Centroids of Fuzzy Numbers", *Fuzzy Sets and Systems*, 157 (2006), 919-926.

An Enhanced Fuzzy Query Approach for Relational Databases

Sabour. A. A.**, Gadallah. A. M.*, and Hefny. H. A.*

Abstract

This paper presents a new approach for building fuzzy databases queries through a new tool called "Fuzzy Query Tool (FQT)". Several fuzzy database models have been proposed. But in these works, fuzziness is introduced at different way. The majority of existing architecture proposals interacts with fuzzy SQL through fuzzy interpreter that converts Fuzzy SQL Queries to Crisp Queries. But in this proposed approach does not follow the idea of the translator, but make Fuzzy SQL built in the database to user call it directly from the database engine. The added value of this work is get a new model makes accelerates the execution of Fuzzy SQL statements than ever before. Next to the ability to write Fuzzy Statements in all clauses of retrieval statement (SELECT) not in the condition clause (WHERE) only as usual.

Keywords : FSQ, Fuzzy Query, SQLf, Fuzzy Logic, Fuzzy Sets, Fuzzy Relational Databases, Fuzzy SQL

1 Introduction

Structured Query Language (SQL) is a very essential tool. It handles data, which is crisp and precise in nature but it is unable to satisfy the needs for data which is uncertain, imprecise, inapplicable and vague in nature. Normal natural language has a lot of vagueness and ambiguity. However, while applying one's thoughts as a query in terms of natural languages into the database, a lot of problems are experienced due to the inefficiency of DBMS to handle such queries. Consider the query "Give me the full names of the Young terrorists who were involved in the recent bomb blast that occurred in the region in and around GUJRAAT". This query cannot be processed directly by the SQL. But it can processed by Fuzzy Query.

Database Management System (DBMS) are extremely useful S/W products which have been used in much kind of systems. Since for a human being the main communication is natural language, this makes it difficult to handle fuzzy, unambiguous and vague information etc. To handle such vague information several works have been developed for the fuzzy query language for fuzzy queries of databases [1].

*Computer Science Dept., Institute of Statistics, Cairo Univ.

**MSc Student, Information Systems Dept., Institute of Statistics, Cairo Univ.

Most of the previously developed fuzzy SQL mechanisms have the following drawbacks:

- a) The inability for dealing with multi subjective views when interacting with multiusers.
- b) The lack of a standardized format in writing statements (no syntax), making it difficult to create the fuzzy query easily.
- c) Most of proposed systems defined a meta base named Fuzzy Meta Knowledge Base (FMB) formed by a set of tables which extend the RDBMS dictionary or catalog in order to store all necessary information to describe and to manipulate fuzzy attributes [2]. This slows down the process of querying, they first inquire about the values of attributes in FMB and then create a fuzzy query with filtering selected values.
- d) Problem of processing nested Fuzzy SQL queries by means of un-nesting. Nested query is an important mechanism in Fuzzy SQL not to mention its efficient execution is a more important issue in Fuzzy SQL than in standard SQL[3].

The rest of this paper is organized as follows: section 2 is an overview about basic concepts of fuzzy set theory. Related works were introduced in section 3. Section 4 presents the proposed fuzzy query approach. The proposed tool and illustrative case study were given in section 5. The conclusion is given in section 6.

2 Basic Concepts of fuzzy set theory

In this section, we first review some basic definitions from fuzzy set theory that will be useful throughout the paper, then define fuzzy query is introduced.

2.1 Fuzzy Sets

Fuzzy set is a set with smooth boundaries. Fuzzy set theory generalizes classical set theory to allow partial membership to the set, and it was developed to cover areas that cannot be covered by classical set theory.

A fuzzy set A in the universe of discourse U is characterized by the membership function μ_A given by μ_A :

$U \rightarrow [0, 1]$ and A is defined as the set of ordered pairs $A = \{(u, \mu_A(u)) : u \in U\}$, where $\mu_A(u)$ is the grade of membership of element u in the set A .

2.2 Membership functions

Fuzzy set is defined as a membership function. Membership function maps an element in the universe of discourse X (i.e. domain of the set) to some value in the interval $[0, 1]$:

$$\mu_A: X \rightarrow [0, 1]$$

The crisp set membership function is given below:

$$\chi_A: X \rightarrow [0, 1]$$

Naturally, we can define three cases of an element's membership:

- If $\mu_A(x) = 1$, then x is a full member of fuzzy set A .
- If $\mu_A(x) = 0$, then x is not a member of fuzzy set A .
- If $\mu_A(x) = c$, $0 < c < 1$, then x is a member of fuzzy set A with degree of membership c .

For finite universe of discourse, one way of listing the members of a fuzzy set A is

$$A = \{\mu_A(x_1)/x_1 + \mu_A(x_2)/x_2 + \dots + \mu_A(x_n)/x_n\}$$

where n = the number of elements in A .

If the universe of discourse is **continuous** and ordered, we can define a fuzzy set by its membership function [7].

Examples of the simplest membership functions are [14]:

- Triangular Membership Function
- Trapezoidal Membership Function
- Gaussian Membership Function
- Bell Shape Membership Function
- Sigmoidal Membership Function
- S Membership Function
- π Membership Function

2.3 Fuzzy Sets Operations

For crisp sets, the basic operations are, namely,

- Union, OR
- Intersection, AND
- Complement, NOT

As an analogy, for fuzzy sets we define fuzzy operators that allow us to manipulate the fuzzy sets. We similarly have fuzzy complements, intersection and union operators but they are not uniquely defined i.e. as membership functions, they are also context - dependent. However an important dissimilarity exists there between traditional set / logic and fuzzy set theory.

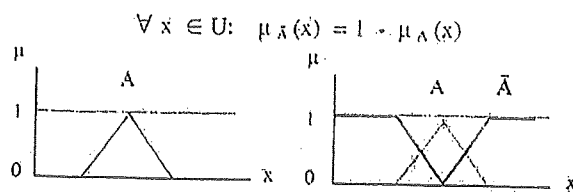
Traditionally there is a distinction between a union operation of sets and "OR" of logic as is the case with intersection and "AND" also. But in fuzzy theory there is no such distinction between the logical and set Operators Fuzzy sets allow operations of union, intersection, and complement [8].

These operations can be used when **linguistic hedges**, such as "very high" or "slightly high" or "Fairly high" ...etc. are used

- Fuzzy union = Fuzzy OR
- Fuzzy intersection = Fuzzy AND
- Fuzzy complement = Fuzzy NOT

We define some standard fuzzy operations as:

- Fuzzy Complement $\sim A(x) = 1 - A(x)$

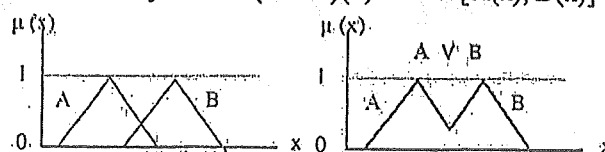


(a)The fuzzy Set A.

(b)The complement of fuzzy Set A.

Figure 1 complement of a fuzzy set.

■ Fuzzy Union $(A \cup B)(x) = \max [A(x), B(x)]$

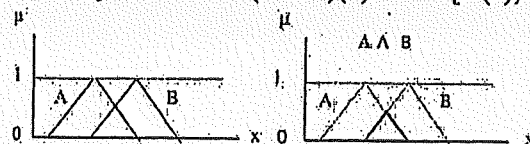


(a) Two fuzzy Sets A and B.

(b) The union of fuzzy Sets A and B.

Figure 2 the union of two fuzzy sets.

■ Fuzzy Intersection $(A \cap B)(x) = \min [A(x), B(x)]$



(a) Two fuzzy Sets A and B.

(b) Intersection of fuzzy Sets A and B.

Figure 3 the intersection of two fuzzy sets.

2.4 Linguistic variable/label and linguistic values

A linguistic variable is a variable having fuzzy (i.e. linguistic) values. For example, the linguistic variable "height" may have the linguistic values "Tall", "Medium" And "short". Each of such fuzzy values is represented by its membership function. The fuzzy vales "Tall", "Medium" and "short" are shown in figure 4.

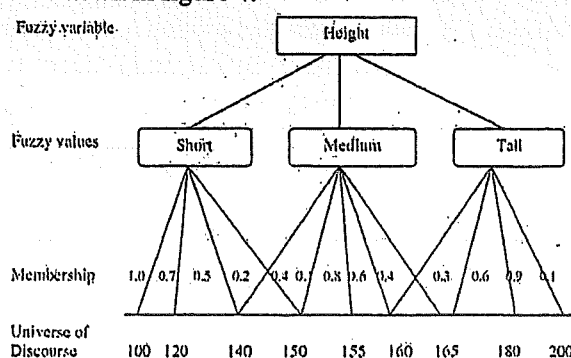


Figure 4 the linguistic variable "Height" and its linguistic values "Tall", "Medium" and "Short"

Four different levels can be distinguished in the definition of a linguistic variable as shown in figure 4. At the top level, we have the name of the variable (e.g., height). At the level below it, we have the labels of fuzzy values (starting with an initial set of values called primary values or term set). Further, down we have membership values, and at the bottom, we have a universe of discourse. It is important to observe that linguistic variables have a dual nature; at higher levels, we have a symbolic linguistic form, and at lower levels, we have a well-defined quantitative analytical form that is, the membership function.

2.5 The Fuzzy SQL Language

Fuzzy SQL is a powerful extension to the Structured Query Language (SQL). Conventional SQL joins tables and screens (or selects) rows based on comparisons using Boolean Logic and ordinary arithmetic.

Fuzzy SQL allows defines semantic concepts on data elements. Thus we can define, for Budget, as an example, what values are Low, Medium, and High, The range of these values can overlap in a natural way, so that a value can be Low to some degree and also Medium to some degree.

Example:[9] "Give me the names of the young age and high salary employees". This query is modelled in SQL language as follows:

```
SELECT Emp_name, Emp_age, Emp_salary
FROM Employee
WHERE (Emp_age between 22 and 30)
      and (Emp_salary>8)
```

The FSQL server uses Database Objects (packages, procedures, function, ... etc.) to model the different types of fuzzy attributes.

The attribute age, presented in Fig. 5, has the linguistic labels Young, Adult and Old, defined on the trapezoidal possibility distributions as following:

YOUNG (18, 22, 30, 35)
ADULT (25, 32, 45, 50)
OLD (50, 55, 62, 70)

An approximate value has a margin of 5.

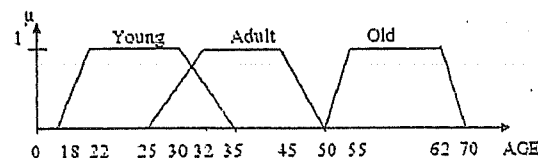


Figure 5 fuzzy membership function for AGE attribute.

The attribute SALARY, presented in Fig. 6, has the linguistic labels Low, Medium and High, defined on the trapezoidal possibility distributions as following:

LOW (1, 1.5, 2.5, 4)
MEDIUM (3, 4.5, 6.5, 8)
HIGH (6, 8, 10, 12)

An approximate value has a margin of 1.5 L.

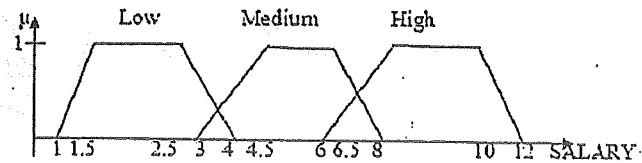


Figure 6 fuzzy membership function for Salary attribute

Query syntax after those Settings will be written as follows in our proposed system:

```
SELECT Emp_name, Emp_age, Emp_salary
FROM Employee
WHERE Age.Young(Emp_age)
      and Salary.High(Emp_salary)
```

3 Previous Works of Fuzzy Querying in Databases

In this section, we present the different latest architectures proposed to model the flexible queries.

3.1 The Fuzzy Relational Database Management System(FRDBMS) Architecture.

The Fuzzy Relational Database Management System(FRDBMS) Architecture Proposed by Grissa A. and Ben Hassine M.[2]

The weak coupling approach with Oracle DBMS. The FRDBMS proposed respects the GEFERD model. The language of description and manipulation of the data is therefore FSQL. Seen that the FSQL language is an extension of the SQL language, a FRDBMS can model a RDB or a FRDB. The principle of this coupling is the definition of a software layer that allows the transformation of the command written by the user in FSQL language in their equivalent written in SQL. This principle is illustrated in the Figure 7.

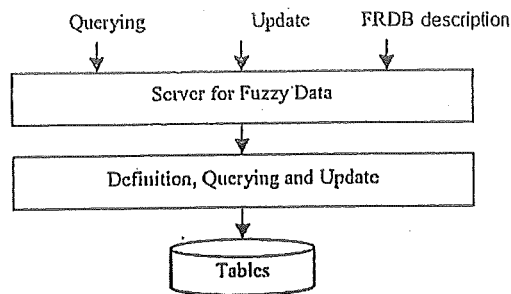


Figure 7 architecture of the proposed FRDBMS model.

The drawbacks of this architecture are:

- The inability for dealing with multi subjective views when interacting with multiusers.
- The lack of a standardized format in writing statements.
- Problem of processing nested Fuzzy SQL queries.
- Problem of Converting Fuzzy Queries to Crisp then Filter the results in order to get valid tuples and that slows down the process of querying.

3.2 The Architecture Proposed by Mishra J [1].

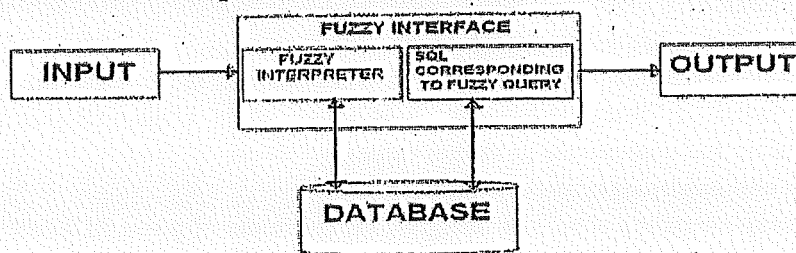


Figure 8 Fuzzy Query Architecture proposed by Mishra J.

Working principle of all components of the above proposed architecture is given below:

- INPUT:** Here submit fuzzy query in natural language form.
- FUZZY INTERFACE:** Fuzzy interface has two components namely FUZZY INTERPRETER and FUZZY SQL
- FUZZY INTERPRETER:** In this phase fuzzy attributes are identified from the given fuzzy query. Based on the fuzzy attribute fuzzy interpreter fetches the domain set of that attribute from the database and then finds the membership value for fuzzy equality of each domain value of the fuzzy attribute with respect to the given fuzzy data in the fuzzy query using the proposed formula. If the query has more than one fuzzy attributes then in similar way it finds the membership value for fuzzy equality for other fuzzy attributes. Finally it takes fuzzy intersection of all membership values which are obtained from different fuzzy attributes to get the membership value for each tuple of the relation.

- **FUZZY SQL:** In this phase decision maker gives the α -cut value. Based on this α -cut value SQL syntax for the given fuzzy query will be generated which we will be called as fuzzy SQL.
- **OUTPUT:** Finally above fuzzy SQL is submitted to the database to get the desired result.

The drawbacks of this architecture are:

- a) The inability for dealing with multi subjective views when interacting with multiusers.
- b) The lack of a standardized format in writing statements.
- c) Problem of processing nested Fuzzy SQL queries.
- d) Problem of Converting Fuzzy Queries to Crisp then Filter the results in order to get valid tuples and that slows down the process of querying.

3.3 The Fuzzy Classification Framework Architecture Proposed by Veryha [6].

This framework is composed of the following steps:

Step-1 Design of DB tables:

Before the composing of queries, DB administrator or DB designer should have designed DB tables to query them.

Step-2 Design of Fuzzy Meta-Knowledgebase:

Meta-Knowledge base means that the additional DB table and activities to support the fuzzy classification based on fuzzy interpreter. Additional tables contain linguistic variables, membership values and descriptions of atomic values.

Step-3 Design and implementation of interpreter:

In this step, fuzzy interpreter should be designed on the frame of given RDBMS using lexical and syntactical analysis of queries. The interpreter transforms the fuzzy SQL into a native SQL. This step should be carried out by software developer or DB programmer.

Step-4 Generation of DB reports and views:

As a result of fuzzy classification by using conventional SQL queries and fuzzy interpreter, fuzzy classified data is presented.

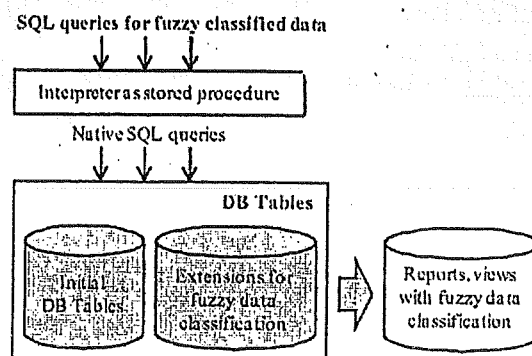


Figure 11 Fuzzy classification framework proposed by Veryha [6].

The drawbacks of this architecture are:

- a) Defined meta-base named Fuzzy Meta Knowledge Base (FMB), This slows down the process of querying.

4 The new proposed architecture

In this section, we present the new proposed architecture to model the flexible fuzzy queries.

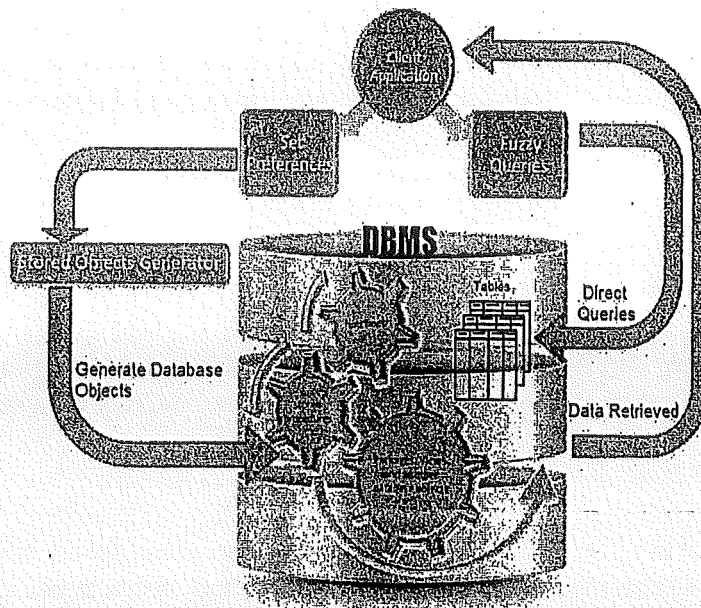


Figure 12 the new Fuzzy query architecture.

The processing scenario of the proposed fuzzy query approach

Working principle of all components of the above proposed architecture is given below:

- User works with a graphical interface to defines the Fuzzy Linguistic Variables and values preferences (linguistic values).
 - Preference SQL It provides foundations for a Pareto-based preference model for database systems; a preference is formulated as a strict partial order on a set of attribute values; It introduces a number of preference operators to express and compose preferences; A compensatory strategy between atomic conditions is not possible due to the fact that Preference SQL makes use of functions for evaluating the distance with which a tuple disagrees with an atomic condition. Moreover, the most preferred tuples are returned to the user without being capable to distinguish how better is one tuple compared to another [4].

- Preferences is user oriented, its work according to the settings sets by each user separately. (User Profiling).
- The program begins by creating PL-SQL Statements that creates database stored objects (such as packages, procedures, functions, views, etc.) within the database, these objects save all data that object require inside them instead of using Fuzzy Meta Knowledge Base (FMB) when running, in order to accelerate execution and lifting performance and efficiency when performing fuzzy queries.
- Then the user can write a Fuzzy Query or generated from query builder GUI, the query is identical to the standard SQL language by using stored objects within the database, the query will go directly to the Database engine to execute directly in one step as it does with regular Crisp SQL statements.

5 Fuzzy Query Tool (FQT)

FQT tool that apply the new proposed approach, and Will be clarified as follows
In this section, we present the new tool FQT that apply the new proposed approach.

5.1 Login to Fuzzy Tool

Main Window which use Login to the system Figure 13. The system specify from this window the user to show his linguistic variables, linguistic values, and enable to run queries with his setting that set before.

Note that the user enters the system and is also user in DBMS and has all the powers identified him by the DBMS. Each user can also be running all its own Fuzzy SQL all databases, which has the right to access and deal with it on the face of view.

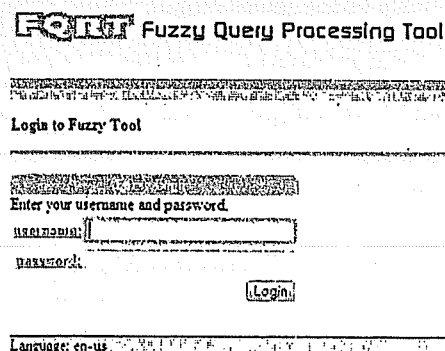


Figure 13 Login Screen

5.2 Main Window

The main screen displays the user basic options for dealing in the Fuzzy SQL environment which will be followed by explaining each of them separately. The next figure-14 show it.

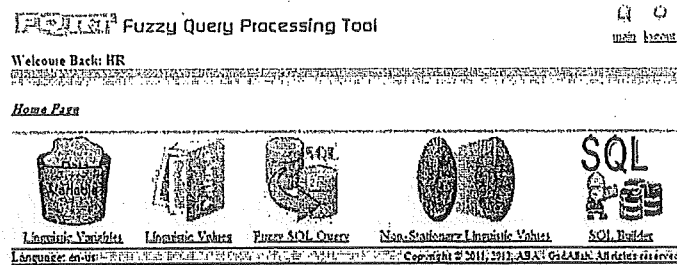


Figure 14 Main Window

5.3 Linguistic Variables

This screen allows create, modify or delete any Linguistic Variables into his personal account, a Write name and then click on the Create or Replace as shown in Figure 15.

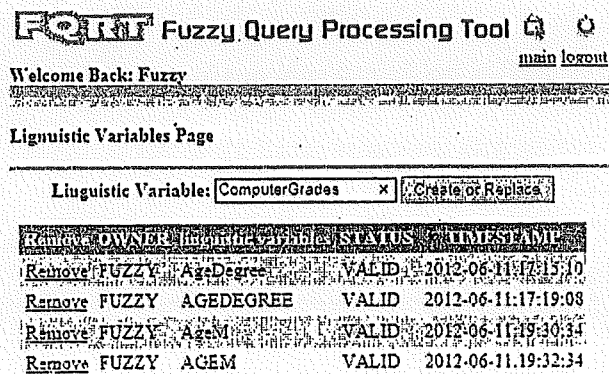
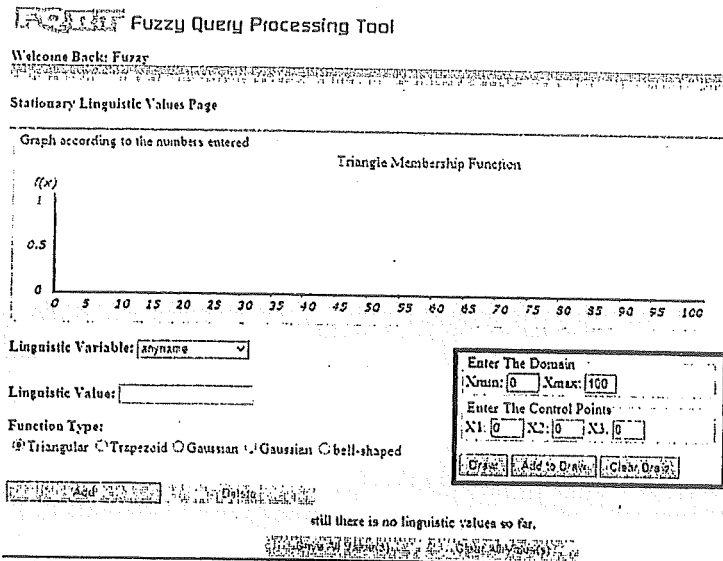


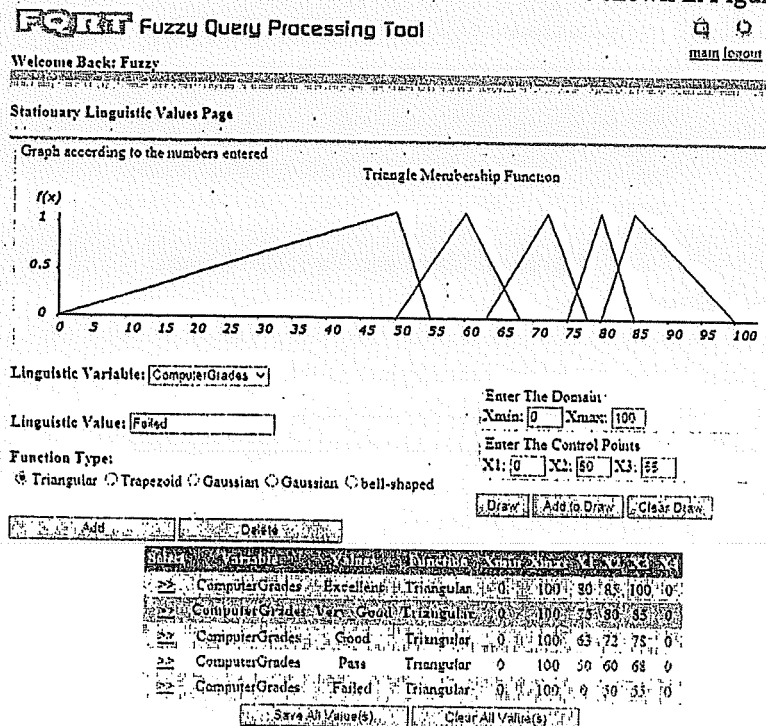
Figure 15 Linguistic Variables Screen

5.4 Linguistic Values

The first to be selected in this screen is the Linguistic Variable. And if the Linguistic Variable contains Linguistic Values set by the user appears to full amendment to allow him if he wanted. If not, then the user selects names of Linguistic value to be added to the Group of Linguistic Variable then determines Function Type to calculate and draw all the Linguistic Values According selected for the function. Then the user selects the domain values that will draw in X-axis their values through lesser and greater value on the drawing. Then the user enters the "Control Points" values that are determined according to the "Function Type" chosen by the user, according to equation implementation. This scenario specified in the next screen figure 16.



The user can draw these values (Control Points) only, or add to the values set by or delete the previous drawing. Then add these values to the Linguistic Value to add other Linguistic Value for the same Linguistic Variable. And so on, until the user saves all the values as shown in Figure 17.



5.5 Fuzzy SQL Query

This screen enables the user to enter sentences the Fuzzy Query then run to display the result as shown in Figure Next. Which shows the result and how to calling and use Linguistic Variable / Values on a numeric field.

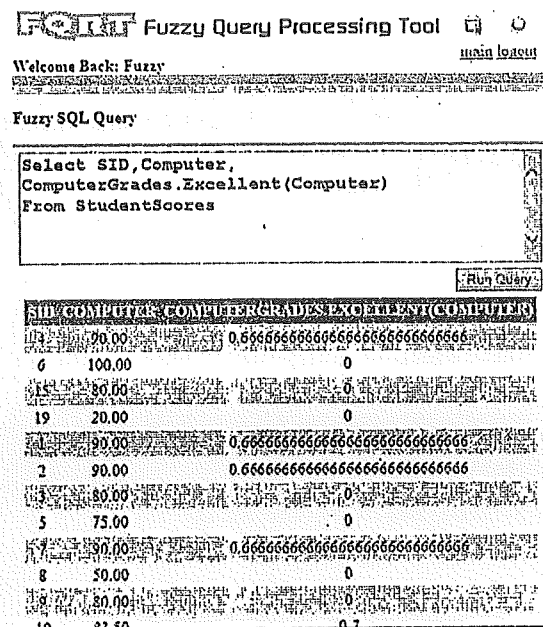


Figure 18 Fuzzy SQL Query

The system support to write Fuzzy Expression as a condition in Where Clause as follows.

```
Select SID, Computer,
ComputerGrades.Excellent(Computer)
From StudentScores
Where ComputerGrades.Excellent(Computer) >0
```

The system support to write Fuzzy Expression into functions or nested functions with calculated field.

```
select sname, round((SysDate-Bdate)/365) age,
round(AgeDegree.OLD((SysDate-Bdate)/365),2)AgeDegree
from student
where AgeDegree.OLD((SysDate-Bdate)/365) >0
```

Run Query

SNAME	AGE	AGEDegree
Abser Amin	44	0.22
Retal Ayman	44	0.22
Kareem Aly	43	0.15
Sameer Abd Allah	41	0.06
Fawzy Lela	41	0.04

Figure 19 Result of fuzzy query uses nested functions

The system support to work with joins or retrieved data from multiple tables, and sort data according to Matching Degree Field.

```
Select S.SID,S.SName,Computer,
Round( ComputerGrades.Good(Computer),3 )
as degree
From Student S, StudentScores SS
Where S.SID = SS.SID
And ComputerGrades.Good(Computer) >0
Order by Degree
```

Run Query

SID	SNAME	COMPUTER	DEGREE
5	Iman Mubhamad	75.00	0.5
18	Samra Ahmad	70.00	0.778
14	Saad Aly Abd Allah	70.00	0.778

Figure 20 Result of fuzzy query uses sort & multiple tables

Support creating a View based on Fuzzy Expressions

```
Create Or replace View AgeDegree_View
as
select round((SysDate-Bdate)/365) age,
round(AgeDegree.OLD((SysDate-Bdate)/365),2)AgeDegree
from student
where AgeDegree.OLD((SysDate-Bdate)/365) >0
```

6 Conclusion

This paper presents a new approach for building fuzzy databases queries through a new tool called "Fuzzy Query Tool (FQT)".Its make Fuzzy SQL built in the database to user call it directly from the database engine. The added value of this work is get a new model makes accelerates the execution of Fuzzy SQL statements than ever before. Next to the ability to write Fuzzy Statements in all clauses of retrieval statement (SELECT) not in the condition clause (WHERE) only as usual.

7 References

- [1] Mishra J., "Fuzzy Query Processing", International Journal of Research and Reviews in Next Generation Networks, Vol. 1, No. 1, March 2011.
- [2] Grissa A. and Ben Hassine M., "New Architecture of Fuzzy Database Management Systems", The International Arab Journal of Information Technology, Vol. 6, No. 3, July 2009.
- [3] Qi Yang et al., "Efficient Processing of Nested Fuzzy SQL Queries in a Fuzzy Database", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 13, NO. 6, NOVEMBER/DECEMBER 2001.
- [4] Abbaci K. et al., "Selecting and Ranking Business Processes with Preferences: An Approach Based on Fuzzy Sets", International Conference on Cooperative Information Systems, Greece (2011)
- [5] Tamani N. et al., "Bipolarity in Flexible Querying of Information Systems Dedicated to Multimodal Transport Networks", The 10th International Symposium on Programming and Systems, hal-00657295, version 1 - 6 Jan 2012.
- [6] Singh K. et al., "Study of Imperfect Information Representation and FSQL processing", International Journal of Scientific & Engineering Research Volume 3, Issue 5, May-2012
- [7] Wibisono O., "Simple Fuzzy SELECT Query on Crisp Database", MakalahIF4058 TopikKhususInformatikaI – Sem. II Tahun 2011/2012 .
- [8] Garg A., and Rishi R., "Querying Capability Enhancement in Database Using Fuzzy Logic", Global Journal of Computer Science and Technology, Vol. 12 Issue 6 Version 1.0 March 2012.
- [9] Jain N. and Shukla S., "Fuzzy Databases Using Extended Fuzzy C-Means Clustering", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, May-Jun 2012.
- [10] Olufade F. and Oladeji, P., "A FUZZY SEARCH MODEL FOR DEALING WITH RETRIEVAL ISSUES IN SOME CLASSES OF DIRTY DATA", Proceedings of the 16th International Conference on Information Quality, 2011
- [11] Kumar A. et al., "Integration of Fuzzy Databases: Problems & Solutions", International Journal of Computer Applications (0975 – 8887), Volume 2 – No.3, May 2010.
- [12] Ashu, "Intuitionistic Fuzzy Approach to handle Imprecise Humanistic Queries in Databases", International Journal of Computer Applications (0975 – 8887), Volume 43– No.20, April 2012.
- [13] Ravi G., SANTHOSH J., "A survey on retrieving Contextual User Profiles from Search Engine Repository", International Journal Of Computational Engineering Research, Vol. 2 Issue. 4, 2012.
- [14] Ir. Wahidin "Fuzzy Control", 2007.
- [15] GadAllah A. and Hefny H. "An Efficient Database Query Processing Tool Based on Fuzzy Logic", The 37th annual conference on statistics, 2002.
- [16] Shawky A. et al. "FRDBM: A Tool For Building Fuzzy Relational Databases", The Egyptian Computer Journal, 2006.

A Survey of Extracting Arabic Multiword Expressions

Samah Meghawry¹ Tarek Elghazaly¹ Abeer El-korany² Akram Salah²

Abstract

Considerable attention has been given to the problem of Multiword Expression (MWE) identification and treatment, for natural language processing(NLP) tasks like parsing and generation, in order to improve the quality of results. Statistical methods have been often employed for MWE identification, such as an inexpensive and language independent way of finding co-occurrence patterns. On the other hand, more linguistically motivated methods for identification, which employ information such as part of speech (POS) filters and lexical alignment between languages, can produce more targeted candidate lists. This paper is an overview of the different types and techniques of extracting MWEs. Concluding to that, MWEs is a multifaceted problem that cannot be solved by one simple solution so the hybrid method is the most accurate approach for extracting MWEs.

Key words: MultiwordExpressions, Natural Language Processing, Information Extraction, Statistical Measures and Word Alignment.

1.Introduction

Multi-Word Expressions (MWEs) can be defined as a semantic issue of a phrase where the meaning of the phrase may not be obtained from its constituents in a straightforward manner[1].

MWE refer also to varioustypes of linguistic units or expressions, includingidioms, noun compounds, named entities, complexverb phrases and any other habitual collocations.MWEs pose a particular challenge in empirical NaturalLanguage Processing (NLP), because they alwayshave idiosyncratic interpretations which cannotbe formulated by directly aggregating the semanticsof their constituents [2].A MWEscan be a multiword unit or a collocation of words that co-occur together statistically more than chance. A MWE is a cover term for different types of collocations which vary in their transparency and fixedness. Identifying

1. Institute of Statistics Studies and researches- Cairo University.

2. Faculty of computer and information- Cairo University.

MWEs and understanding their meaning is considered essential to language understanding and of crucial importance for any NLP applications that aim at handling robust language meaning and use.

MWEs are lexical items that consist of multiple orthographic words (e.g., ad hoc, by and large, New York, kick the bucket). MWEs are numerous and constitute a significant portion of the lexicon of any natural language. They are a heterogeneous class of constructions with diverse sets of characteristics, distinguished by their idiosyncratic behavior. Morphologically, some MWEs allow some of their constituents to freely inflect while restricting (or preventing) the inflection of other constituents.

In some cases MWEs may allow constituents to undergo non-standard morphological inflections that they would not undergo in isolation. Syntactically, some MWEs behave like words while others are phrases; some occur in one rigid pattern (and a fixed order), while others permit various syntactic transformations. Semantically, the compositionality of MWEs is gradual, ranging from fully compositional to idiomatic [3].

To be considered as a multi-word term, a sequence of words should fulfill syntactic and semantic conditions. In fact, it can be distinguished to many types of MWEs [4]:

- Idioms (e.g. "down the drain")
- Phrasal verbs (e.g. "rely on")
- Verbs with particles (e.g. "give up")
- Compound nouns (e.g. "book cover")
- Collocations (e.g. "do a favor")

2. Classification of MWE

2.1) Compositional vs. Non-Compositional MWEs

Semantic compositionality, sometimes termed decomposability, is "a means of describing how the overall sense of a given idiom is related to its parts" [5]. An illustrative example of non-compositionality is the expression kick the bucket, where the meaning "die" has no relation to any word in the expression. An example of compositional expressions is the compound noun book cover, where the meaning is directly related to the component parts.

Unfortunately, the assignment of a plus/minus feature of compositionality to an expression is sometimes very elusive.

Most of the time “One cannot really make a binary distinction between compositional and non-compositional MWEs”[6]. They occupy a continuum in a large scale. At one end of the scale there are those expressions that are highly opaque and non-compositional, where the meaning is not traceable to any of the component parts, such as kick the bucket. In the middle of the scale there are those where one or more words are used in an idiosyncratic sense, or use “semantics unavailable outside the MWE” [7], such as spill the beans, which means “to disclose a secret”. At the other end of the scale there are those which are highly compositional, such as book cover, traffic light, health crisis and party meeting.

Non-compositional expressions, or, more accurately, expressions that show any degree of non-compositionality, are termed lexicalized and are automatically eligible to be considered as MWEs. However, in order for compositional expressions to be included in an MWE lexicon, they need to be conventionalized or institutionalized. This means that these expressions have come to such a frequent use that they block the use of other synonyms and near synonyms [8]. When words co-occur in a statistically meaningful way like this they are called collocations. This way, expressions such as book cover and traffic light can be safely added to an MWE lexicon, while health crisis and party meeting cannot.

Flexible vs. Inflexible MWEs

With regard to syntactic and morphological flexibility, MWEs are classified into three types: fixed, semi-fixed and syntactically flexible expressions [9][10][11].

2.2.1) Fixed Expressions

These expressions are lexically, syntactically and morphologically rigid. An expression of this type is considered as a word with spaces (a single word that happens to contain spaces), such as San Francisco and in a nutshell. Some expressions are frozen at the level of the sentence, sometimes termed “frozen texts” [12]. These include proverbs such as Buy cheap, buy twice and pragmatically fixed expressions such as Good morning.

2.2.2) Semi-Fixed Expressions

These expressions can undergo morphological and lexical variations, but still the components of the expression are adjacent. They cannot be reordered or separated by external elements. The variations that can affect semi-fixed expressions are of two types: 1. Morphological variations that express person, number, tense, gender, etc., such as traffic light/lights and kick/kicks/kicked the bucket. 2. Lexical variations. This is the case when a position in the expression is filled by a choice from the set of reflexive pronouns (e.g. prostrate himself/herself), or when one word can be replaced by another (e.g. to sweep something under the carpet/rug).

2.2.3) Syntactically Flexible Expressions

These are the expressions that can either undergo reordering, such as passivization (e.g. the cat was let out of the bag), or allow external elements to intervene between the components (e.g. slow the car down). Here the adjacency of the MWE is disrupted[13].

3. Techniques

Extraction of MWE in general can be classified into two types of MWT extraction approaches: i) statistical approaches and ii) linguistic approaches. Statistical approaches are based on association measures based on frequency and co-occurrence such as T-score, log-likelihood ratio (LLR) [14], FLR and Mutual Information (MI3).

Linguistic approaches make use of morphological, syntactic or semantic information implemented in language-specific rules or programs. Linguistic approaches can be divided into two categories:

Approaches based on morpho-syntactic patterns: Candidate terms selection is based on the internal structure of MWTs. Conditions about the internal structure of terms are encoded as morpho-syntactic patterns. As examples of tools based on this approach we can cite ACABIT [15], Nomino[16]and OntoLearn.

Approaches based on MWEs boundaries detection: locate configurations that cannot be components of MWEs such as punctuations and conjugated verbs. This approach is implemented in Lexter[17].

Table1:A comparison between MWE's extraction techniques

	Statistical	Alignment	Symbolic/ linguistic
Methodology	MWEs extraction are based on lexical association measures – mathematical formulas determining the strength of association between two or more words based on their occurrences and co-occurrences in a text corpus. The higher the association between words, the better chance they form a MWEs.	The automatic word alignment of two parallel texts—a text written in one (Source) language and its translation to another (target) language—tries to identify for each word in a source sentence equivalences in the parallel target sentence.	Linguistic approaches make use of morphological, syntactic or semantic information implemented in Language-specific rules or programs.
Advantage	They can be used either for ranking (candidates with high scores at the top) or for classification (by setting a threshold and discarding all bigrams below this threshold)	Word alignment is able to attach semantic information to word and multiword units, by means of their target language counterparts. This approach can help to considerably reduce and accelerate lexicographic work by generating lists of MWEs with their translations, for the construction of bilingual resources, and/or with some semantic information for monolingual resources.	Linguistic analysis has more than one level of analysis such as morphological analysis, lexical analysis, syntactic analysis (parsing) and semantic analysis. The focus of corpus analysis is empirical, whereas the interpretation can be either qualitative or quantitative.
limitation	Due to corpus size, statistical measures have mostly been applied to bigrams and trigrams, and it becomes more problematic to extract MWEs of more than three words.	Statistical measures or linguistic rules should be support the alignment process to get accurate results in extraction process	It is difficult to apply symbolic methods to data with no syntactic annotations. Pure linguistic approaches are language dependent and not flexible enough to cope with complex structures of MWTs

3.1) Statistical Technique

Statistical measures of association have been widely employed in the identification of MWEs. The idea behind their use is that they are an inexpensive language and type independent means of detecting recurrent patterns. If a group of words co-occurs with significantly high frequency when compared to the frequencies of the individual words, then they may form an MWE. Indeed, measures such as Pointwise Mutual Information (PMI), Mutual Information (MI), log-likelihood [18] and others have been employed for this task, and some of them seem to provide more accurate predictions of MWEness than others. In fact, in a comparison of some measures for the type-independent detection of MWEs, MI seemed to differentiate MWEs from non-MWEs, but the same was not true of χ^2 [19].

3.2) Alignment Technique

The automatic word alignment of two parallel texts — a text written in one (source) language and its translation to another (target) language— is the process of searching for correspondences between source and target words and sequences of words. For each word in a source sentence equivalences in the parallel target sentence are looked for. Therefore, taking into account a word alignment between a source word sequence S ($S = s_1 \dots s_n$ with $n \geq 2$) and a target word sequence T ($T = t_1 \dots t_m$ with $m \geq 1$), that is $S \leftrightarrow T$, the alignment-based MWE extraction method assumes that: (a) S and T share some semantic features, and (b) S may be a MWE. In other words, the alignment-based MWE extraction method states that the sequence S will be a MWE candidate if it is aligned with a sequence T composed of one or more words (a $n : m$ alignment with $n \geq 2$ and $m \geq 1$).

The alignment-based MWE extraction method does not rely on the conceptual asymmetries between languages since it does not expect that a source sequence of words be aligned with a single target word. The method looks for the sequences of source words that are frequently joined together during the alignment despite the number of target words involved. These features indicate that the method prioritizes precision in spite of recall.

It is also important to say that although the sequences of source and target words resemble the phrases used in the phrase-based statistical machine translation (SMT), they are indeed a refinement of them. More specifically, although both approaches rely on word alignments performed by GIZA++ [20], in the alignment-based approach not all sequences of words are

considered as phrases (and MWE candidates) but just those with an alignment $n : m$ ($n \geq 2$) with a target sequence[21].

3.3) Symbolic Approach

Symbolic approach- it can be also called Corpus-based Linguistic analysis- is used to identify linguistic markers which point to certain morphological, syntactic or semantic relations between MWEs.

This approach has the advantage of computational tractability but is inherently limited to uncovering only relations explicitly identified through the targeted lexico-syntactic patterns [22].

In this approach, morphotactic (rules governing the combination of morphemes, which are meaning bearing units in the language) and orthographic (spelling rules) rules are programmed into a finite state transducer (FST). Koskeniemi proposed a two-level system for language morphology, which led to Antworth's two-level morphology system PCKIMMO [23][24]. Later, Beesley and Buckwalter developed an Arabic morphology system, ALPNET that uses a slightly enhanced implementation of PC-KIMMO [25]. Currently, ALPNET is owned by Xerox and uses Xerox Finite-State Morphology tools [26]. However, this approach was developed [27] for requiring excessive manual processing to state rules in an FST and for the ability only to analyze words that appear in Arabic dictionaries. Kiraz summarized many variations of the FST approach [28]. Much information on two-level morphology and PC-Kimmo is available in the PC-KIMMO user's guide [29] [30].

4. Related work

Arabic is highly ambiguous. Some recent works which dealt with this problem use either pure linguistic or hybrid approaches. For example, [31] presented a pure linguistic approach for handling Arabic MWEs. It is based on a lexicon of MWEs constructed manually. Then the system tries to identify other variations using a morphological analyzer, a white space normalizer and a tokenizer. Precise rules allow taking into account morphological features such as gender and definiteness to extract MWEs. The MWEs structures are described as trees that can be parsed to identify the role of each constituent. However some types of MWEs are ignored such as substitution compound nouns. Besides on, the relevance of the extracted candidates is not computed because the lack of statistical measures.

Boulaknadel[32] presented a hybrid approach to extract MWEs from Arabic documents. They used patterns based on POS to identify candidate terms. Then, candidate terms are ranked using four statistical measures: MI3, T-scores, LLR and FLR. They reported that LLR is the best measure as far as precision is concerned. In fact, many critics can be addressed to this approach. First, the approach does not include a morphological analysis step. The used POS tagger [33] is unable to separate affixes, conjunctions and some prepositions from nouns and adjectives. The lack of a morphological analysis step obliged the authors to identify in second step- variants of the already identified MWE. Thus, they identify graphical variants, inflectional variants, morpho-syntactic and syntactic variants. Second, POS tagging does not allow taking into account many features while defining MWE patterns. For example, we cannot impose constraints about the gender and/or the number of the MWE constituents. Third, this approach does not deal with syntactic ambiguities. Let consider the following example: “الإنهاء من بناء الدار” (the end of house building).

Analyzing this example will lead to two different parse trees. The MWE extraction process should allow identifying the different parse trees and select the correct one.

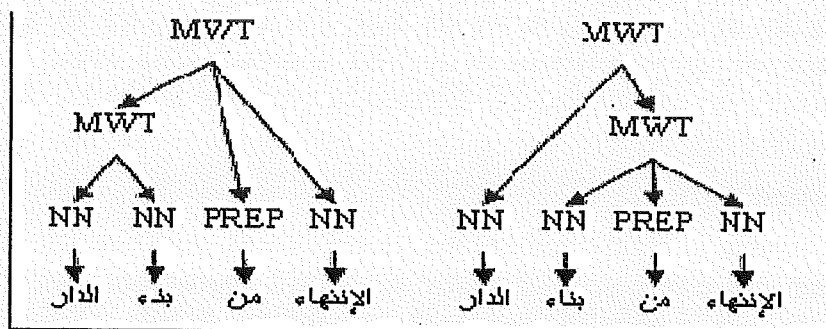


Fig1: Example of ambiguities in MWE extraction.

Fourth, this approach does not allow recognizing the internal structure of MWTs. As we have previously shown, the Arabic language defines different roles of MWT constituents.

Mohamed Attia and others tried to extract Arabic multi-word Expression; using three complementary approaches applicable for three different data sources.

Attia extracted MWEs from heterogeneous data resources using three approaches: (a) crosslingual correspondence asymmetries which relied on the many-to-one relations in inter-wiki links, (b) translation-based extraction, which employs the automatic translation of PWN.MWEs into Arabic and uses different search engines to filter the translation output, and (c) corpus-based statistics, which applies lexical association measures to detect habitual collocations in a large un-annotated corpus. The focus was on nominal MWEs because the vast majority of MWEs are nouns

Attia concluded that the identification of MWEs is too complex to be dealt with by one simple solution. The choice of approach depends, to a large extent, on the type of data resources used. As Arabic has a rich and complex morphology, lemmatizing the text to reduce inflectional forms. These approaches prove to be a fruitful ground for large scale extraction of Arabic MWEs.

5) Conclusion

A considerable amount of research has focused on the identification and extraction of MWEs. Given the heterogeneity of MWEs, different approaches were devised. Broadly speaking, work on the extraction of MWEs revolves around four approaches: (a) statistical methods which use association measures to rank MWE candidates [34]; (b) symbolic methods which use morpho-syntactic patterns [35] ; (c) hybrid methods which use both statistical measures and linguistic filters [36] [37] and (d) word alignment [38].

The identification of MWEs is too complex to be dealt with by one simple solution. The choice of approach depends, to a large extent, on the type of data resources used. Set of steps should be taken for accurately extracting MWEs in general represented in the following:

1. Determining which class to deal with also there is who determine a subclass of the chosen class to extract.
2. Studying the properties of chosen class well.
3. Studying the MWE extraction techniques well to determine which technique is appropriate or if there is a need to a hybrid technique.
4. Determining the resources that we will deal with, it is better choose a resource or resources that others worked on to compare with their results.
5. Preprocessing the resource.
6. Applying the chosen technique.

References

1. T. Chakraborty, D. Das and S.iBandyopadhyay, Semantic Clustering: an Attempt to Identify Multiword Expressions in Bengali Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 8–13,Portland, Oregon.
2. Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger Lecture Notes in Computer Science, 2002, Volume 2276, Computational Linguistics and Intelligent Text Processing, Pages 189-206.
3. Bannard, Colin, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In Francis Bond, Anna Korhonen, Diana McCarthy and Aline Villavicencio, editors, Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, pages 65–72.
4. BOUNHAS, Y. SLIMANI, A hybrid Approach for Arabic Multi-Word Term Extraction, 978-1-4244-4538-7/09/\$25.00 ©2009 IEEE.
5. Sag, Ivan A., Baldwin, Timothy, Bond, Francis, Copestake, Ann, and Flickinger, Dan. 2001. Multi-word Expressions: A Pain in the Neck for NLP. Paper presented at LinGO Working Papers, Stanford University, CA.
6. Venkatapathy, Sriram. 2004. Overview of my work on Multi-word expressions and Semantic Role Labeling.
7. Baldwin, Timothy, Bannard, Colin, Tanaka, Takaaki, and Widdows, Dominic. 2003. An Empirical Model of Multiword Expression Decomposability. Paper presented at the ACL- 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment, Sapporo, Japan.
8. Nerima, Luka, Seretan, Violeta, and Wehrli, Eric. 2003. Creating a Multilingual Collocations Dictionary from Large Text Corpora. Paper presented at 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL03), Budapest, Hungary.
9. Baldwin, Timothy. 2004. Multiword Expressions, an Advanced Course. Paper presented at The Australasian Language Technology Summer School (ALTSS 2004), Sydney, Australia.
10. Oflazer, Kemal, Uglu, ?zlem ?etino, and Say, Bilge. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. Paper presented at Second ACL Workshop on Multiword Expressions: Integrating Processing, Spain.
11. Sag, Ivan A., Baldwin, Timothy, Bond, Francis, Copestake, Ann, and Flickinger, Dan. 2001. Multi-word Expressions: A Pain in the Neck for NLP. Paper presented at LinGO Working Papers, Stanford University, CA.
12. Guenthner, Frantz, and Blanco, Xavier. 2004. Multi-Lexemic Expressions: an overview. In Lexique, Syntaxe et Lexique-Grammaire, eds. Christian Leclère,

- TricLaporte, MireillePiot and Max Silberztein. Philadelphia PA, USA: John Benjamins.
13. Attia, Mohammed. 2006. Accommodating Multiword Expressions in an Arabic LFG Grammar. In Salakoski, Tapio, FilipGinter, SampoPyysalo, TapioPahikkala (Eds.): *Advances in Natural Language Processing*. Vol. 4139, pp. 87–98. Springer- Verlag: Berlin, Heidelberg.
 14. T. Dunning, *Accurate Methods for the Statistics of Surprise and Coincidence*, Computational Linguistics, vol. 19(1), pp. 61-74, 1994.
 15. B. Daille, *Approchemixte pour l'extraction de terminologie :statistiquelexicale et filtreslinguistiques*, doctoral thesis, University of Paris 7, 1994.
 16. S. David and P. Plante, *De la nécessité'd'uneapprochemorpho-syntaxique en analyse de textes*, ICO, Intelligence artificielle et sciences cognitives au Québec, vol. 2(3), pp. 140-151, September 1990.
 17. D. Bourigault, *LEXTER, unlogicield'EXtraction de TERminologie, Application à l'acquisition des connaissances à partir de textes*, doctoral thesis, Ecole des Hautesétudes en Sciences Sociales, Paris, 1994.
 18. H. Press William, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical Recipes in C: The Art of Scientific Computing*. Second edition. Cambridge University Press.
 19. Aline Villavicencio, ValiaKordoni, Yi Zhang, Marco Idiart, and Carlos Ramisch. 2007. Validation and Evaluation of Automatically Acquired Multiword Expressions for Grammar Engineering. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1034–1043, Prague, June.
 20. Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 440–447, Hong Kong, China, October.
 21. Helena de Medeiros Caseli, Aline Villavicencio, Andr'e Machado and Maria Jos'e Finatto, 2009, *Statistically-Driven Alignment-Based Multiword Expression Identification for Technical Domains*, *Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, pages 1–8, Suntec, Singapore, 6 August 2009.
 22. Eric SanJuan, James Dowdall, Fidelialbekwe-SanJuan, and Fabio Rinaldi, 2004. A symbolic approach to automatic multiword term structuring *Computer Speech and Language*. Submitted.
 23. Koskenniemi, Kimmo, 1983 "Two Level Morphology: A General Computational Model for Word-form Recognition and Production. Publication No. 11, Dept. of General Linguistics, University of Helsinki, Helsinki.

24. Antworth, Evan L, 1990 PC-KIMMO: a two-level processor for morphological analysis. Occasional Publications in Academic Computing No. 16. Dallas, TX: Summer Institute of Linguistics.
25. Beesley, Kenneth, 1996 Arabic Finite-State Morphological Analysis and Generation. COLING-96, 1996.
26. Ahmed, Mohamed Attia, 2000 A Large-Scale Computational Processor of the Arabic Morphology, and Applications. A Master's Thesis, Faculty of Engineering, Cairo University, Cairo, Egypt.
27. Kiraz, George, 1998 "Arabic Computation Morphology in the West. Proceedings of the 6th International Conference and Exhibition on Multilingual Computing, Cambridge.
28. Antworth, Evan L , 1995 PC-KIMMO User's Guide. www.sil.org/pckimmo/v2/doc/guide.html.
29. Darwish K., 2002 , Building a Shallow Morphological Analyzer in One Day, In Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA.
30. BOUNHAS, Y. SLIMANI, A hybrid Approach for Arabic Multi-Word Term Extraction, 978-1-4244-4538-7/09/\$25.00 ©2009 IEEE.
31. S. Boulaknadel, B. Daille and D. Aboutajdine, A multi-word term extraction program for Arabic language, the 6th international Conference on Language Resources and Evaluation LREC 2008, Marrakech, Morocco, 28-30 1485-1488 May 2008, pp.
32. M. Diab, K. Hacioglu and D. Jurafsky, Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks, in the 5th Meeting of the North American Chapter of the Association for Computational Linguistics/Human Language Technologies Conference (HLT-NAACL04), Boston, Massachusetts, May 2-7 -2004.
33. Van de Cruys, Tim and Bego~na Villada Moir'on. 2006. Lexico-Semantic Multiword Expression Extraction. In P. Dirix et al. (eds.), Computational Linguistics in the Netherlands 2006, pp. 175-190.
34. Vintar, Spela and Darja Fi ser. 2008. Harvesting Multi-Word Expressions from Parallel Corpora. In the 6th International Conference on Language Resources and Evaluation (LREC 2008). Marrakech, Morocco.
35. Boulaknadel, Siham, Beatrice Daille, Driss Aboutajdine. 2009. A multi-word term extraction program for Arabic language In the 6th International Conference on Language Resources and Evaluation (LREC 2008), pp. 630-634, Marrakech, Morocco.
36. Pacific-Asia Conference on Knowledge Discovery and Data (PAKDD 2009), pp. 541-547. Bangkok, Thailand.

37. Moir' on, Bego~na Villada and J'org Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In the Workshop on Multiword Expressions in a Multilingual Context, the 11th Conference of the European Association of Computational Linguistics (EACL 2006), pp. 33–40. Trento, Italy.
38. Y. Tu and D. Roth, Learning English Light Verb Constructions: Contextual or Statistical Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 31-39, Portland, Oregon.
39. M. Carpuat and M. Diab, Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation, Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 242–245, Los Angeles, California
40. Y. Tsvetkov and S. Wintner, Extraction of Multi-word Expressions from Small Parallel Corpora, Coling 2010: Poster Volume, pages 1256–1264, Beijing.
41. M. A. Attia, Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation, doctoral thesis, University of Manchester, Faculty of Humanities, 2008
42. V. Malaisé, Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels, doctoral thesis, University of Paris 7 – Denis Diderot, 2005.
43. Pacific-Asia Conference on Knowledge Discovery and Data (PAKDD 2009), pp. 541–547. Bangkok, Thailand.

1. The first part of the document is a letter from the President of the United States to the Congress, dated January 1, 1801. It is a very important document, as it is the first official communication of the new administration.

2. The second part of the document is a report from the Secretary of the Treasury, dated January 1, 1801. It contains information about the state of the nation's finances.

3. The third part of the document is a report from the Secretary of the Navy, dated January 1, 1801. It contains information about the state of the nation's naval forces.

4. The fourth part of the document is a report from the Secretary of the War, dated January 1, 1801. It contains information about the state of the nation's military forces.

5. The fifth part of the document is a report from the Secretary of the Interior, dated January 1, 1801. It contains information about the state of the nation's internal affairs.

6. The sixth part of the document is a report from the Secretary of the State, dated January 1, 1801. It contains information about the state of the nation's foreign relations.

7. The seventh part of the document is a report from the Secretary of the War, dated January 1, 1801. It contains information about the state of the nation's military forces.

8. The eighth part of the document is a report from the Secretary of the Interior, dated January 1, 1801. It contains information about the state of the nation's internal affairs.

9. The ninth part of the document is a report from the Secretary of the State, dated January 1, 1801. It contains information about the state of the nation's foreign relations.

10. The tenth part of the document is a report from the Secretary of the War, dated January 1, 1801. It contains information about the state of the nation's military forces.

11. The eleventh part of the document is a report from the Secretary of the Interior, dated January 1, 1801. It contains information about the state of the nation's internal affairs.

12. The twelfth part of the document is a report from the Secretary of the State, dated January 1, 1801. It contains information about the state of the nation's foreign relations.

13. The thirteenth part of the document is a report from the Secretary of the War, dated January 1, 1801. It contains information about the state of the nation's military forces.

14. The fourteenth part of the document is a report from the Secretary of the Interior, dated January 1, 1801. It contains information about the state of the nation's internal affairs.

A Gaussian Fuzzy ANP Model to Rank Different Scenarios of Electrical Power Generation in Egypt

Hesham A. Hefny*, H. M. Elsayed** and H. F. Aly***

Abstract

The analytic hierarchy process (AHP) method assumes that the factors presented in the hierarchical structure are independent; however, this assumption may be inappropriate. Therefore AHP technique does not allow for measurement of the possible dependencies among the factors. The analytic network process (ANP) allows for complex interrelationships among decision levels and attributes. In this paper we will develop an efficient fuzzy ANP model which helps decision makers to rank the alternatives for the scenarios of electrical power generation in Egypt.

Keywords: Fuzzy; Analytic network process; Gaussian function; Decision-making.

1. Introduction

Fuzzy rule-based systems have been widely used in a variety of engineering areas such as data mining, pattern recognition, system identification, and process control [Chen et al., 2001]. Fuzzy logic is a key tool to express knowledge of domain experts so that valuable experience of human beings can be incorporated into controllers design and applied to handle real-life situations that the classical control approach finds difficult or impossible to tackle [Wu et al., 2001].

The analytic network process (ANP) is used for tackling multi-attribute decision-making problems in real situations when there is interrelation among decision criteria or alternatives. In the traditional formulation of the ANP, human's judgments are represented as exact numbers. However, in many practical cases the human preference model is uncertain and decision makers might be unable to assign exact numerical values to the comparison judgments. Since some of the evaluation criteria are subjective and qualitative in nature, it is very difficult for the decision-maker to express the preferences using exact numerical values and to provide exact pair-wise comparison judgments. It is more desirable for him to use interval or fuzzy evaluations [Chan et al., 2005]. To improve the ANP method, this paper discusses a fuzzy ANP approach using Gaussian fuzzy numbers to represent decision makers' comparison judgments and extent analysis method to decide the final priority of different decision criteria. The proposed model uses the linguistic variables and Gaussian fuzzy numbers as a pair-wise comparison scale for deriving the priorities of different selection attributes and sub-attributes. In the last step, the priority weights for main attributes, sub-attributes and alternatives are combined to determine the priority weights of the alternatives. The alternative with the highest priority weight is selected as the best alternative. This paper aims to propose a fuzzy ANP

* Institute of Statistical Studies and Research

** Central Agency for Public Mobilization and Statistics

*** Atomic Energy Authority

decision-making support system that helps decision-makers of any Authority in selecting the best alternatives among several offers. Such a kind of systems often requires highly experienced decision makers to consider vague and uncertain information.

Fuzzy set theory offers a possible means of managing these kinds of data or information. On the other hand, ANP offers a means for dealing with different preferences made to different decision alternatives.

Egypt had installed generating capacity of 20 gigawatts (GW) as of 2010, with plans to add 25 GW of additional generating capacity by 2020. Around 90 percent of Egypt's electric generating capacity is thermal (natural gas), with the remaining 10 percent hydroelectric, mostly from the Aswan High Dam. All oil-fired plants have been converted to run on natural gas as their primary fuel. Egypt is also planning to build a part-solar power plant at Kureimat, which will have a total planned capacity of 150 MW. A Netherlands-funded project is building 60 MW of wind power units in the Suez Canal area. Egypt also has a 22-MW nuclear research reactor at Inshas in the Nile Delta, built by INVAP S.A. of Argentina, which began operation in 1997 [MBendi, 2010].

The remainder of the paper is organized as follows: In section 2, an overview of the different methods of electrical power generation is given. In section 3, Analytic network process (ANP) is illustrated. FANP based approach is discussed in section 4. Gaussian fuzzy ANP (GFANP) proposed model to select the best alternative is developed and the steps of each stage of the procedure are explained in detail in section 5. In section 6, results which are produced from the model are discussed, and the paper ends with concluding remarks in section 7.

2. Electrical Power Generation

Electricity generation is the process of generating electric energy from other forms of energy. The fundamental principles of electricity generation were discovered during the 1820s and early 1830s by the British scientist Michael Faraday. His basic method is still used today. Electricity is generated by the movement of a loop of wire, or disc of copper between the poles of a magnet. For electric utilities, it is the first process in the delivery of electricity to consumers [ELECG, 2010]. Each system of the electricity generation has advantages and disadvantages, but many of them pose environmental concerns. Along with continuous global warming, the environmental problems, besides the economic objective, are expected to play more and more important role in the operation of power systems [Zhang et al., 2012]. The cost of electricity generated by different sources is a calculation of the cost of generating electricity at the point of connection to a load or electricity grid. It includes the initial capital, discount rate, as well as the costs of continuous operation, fuel, and maintenance. This type of calculation assists policy makers, researchers and others to guide discussions and decision making [Sahu et al., 2009].

2.1 Thermal Power Stations

A thermal power station is a power plant in which the prime mover is steam driven. It comprises all of the equipment and systems required to produce electricity by using a steam generating boiler fired with fossil fuels or bio-fuels to drive an electrical generator [Ha et al., 2011]. Such power stations are most usually constructed on a very large scale and designed for continuous operation. Water is heated, turns into steam and spins a

steam turbine which drives an electrical generator. After it passes through the turbine, the steam is condensed in a condenser and recycled to where it was heated; this is known as a Rankine cycle. The greatest variation in the design of thermal power stations is due to the different fuel sources [Jiang et al., 2008].

2.1.1 Fossil-Fuel Power Stations

Most electricity today is generated by burning fossil fuels and producing steam which is then used to drive a steam turbine that, in turn, drives an electrical generator. Such systems allow electricity to be generated where it is needed, since fossil fuels can readily be transported. They also take advantage of a large infrastructure designed to support consumer automobiles [Smoot et al., 2003]. The world's supply of fossil fuels is large, but finite. Exhaustion of low-cost fossil fuels will have significant consequences for energy sources as well as for the manufacture of plastics and many other things. Various estimates have been calculated for exactly when it will be exhausted. New sources of fossil fuels keep being discovered, although the rate of discovery is slowing while the difficulty of extraction simultaneously increases [Popov, 2011].

2.1.2 Nuclear Power Stations

Nuclear power plants do not burn fossil fuels and so do not directly emit carbon dioxide; because of the high energy yield of nuclear fuels, the carbon dioxide emitted during mining, enrichment, fabrication and transport of fuel is small when compared with the carbon dioxide emitted by fossil fuels of similar energy yield. A large nuclear power plant may reject waste heat to a natural body of water; this can result in undesirable increase of the water temperature with adverse effect on aquatic life [Visschers et al., 2011]. Emission of radioactivity from a nuclear plant is controlled by regulations. Abnormal operation may result in release of radioactive material on scales ranging from minor to severe, although these scenarios are very rare [Gao et al., 2011].

2.2 Hydroelectric Power Stations

Hydroelectricity is the term referring to electricity generated by hydropower; the production of electrical power through the use of the gravitational force of falling or flowing water. It is the most widely used form of renewable energy, accounting for 16 percent of global electricity consumption, and 3,427 terawatt-hours of electricity production in 2010, which continues the rapid rate of increase experienced between 2003 and 2009 [HDEL, 2011]. Hydropower is produced in 150 countries, with the Asia-Pacific region generating 32 percent of global hydropower in 2010. China is the largest hydroelectricity producer, with 721 terawatt-hours of production in 2010, representing around 17 percent of domestic electricity use. There are now three hydroelectricity plants larger than 10 GW: the Three Gorges Dam in China, Itaipu Dam in Brazil, and Guri Dam in Venezuela [Sovacool et al., 2011].

2.3 Wind Power Stations

Wind power is the conversion of wind energy into a useful form of energy, such as using wind turbines to make electricity, windmills for mechanical power, wind pumps for water pumping or drainage, or sails to propel ships [WINEG, 2010].

A large wind farm may consist of several hundred individual wind turbines which are connected to the electric power transmission network. Offshore wind farms can harness more frequent and powerful winds than are available to land-based installations and have less visual impact on the landscape but construction costs are considerably higher. Small onshore wind facilities are used to provide electricity to isolated locations and utility companies increasingly buy back surplus electricity produced by small domestic wind turbines. Wind power harnesses mechanical energy from the constant flow of air over the surface of the earth [WINEG, 2010].

2.4 Solar Power Stations

Solar energy, radiant light and heat from the sun, has been harnessed by humans since ancient times using a range of ever-modern technologies. Solar energy technologies include solar heating, solar photovoltaic, solar thermal electricity and solar architecture, which can make considerable contributions to solving some of the most urgent problems the world now faces [SOLEG, 2010]. Solar technologies are broadly characterized as either passive solar or active solar depending on the way they capture, convert and distribute solar energy. Active solar techniques include the use of photovoltaic panels and solar thermal collectors to harness the energy. Passive solar techniques include orienting a building to the sun,

selecting materials with favorable thermal mass or light dispersing properties, and designing spaces that naturally circulate air [Xu et al., 2011].

3. Analytic Network Process (ANP) Approach

ANP is a multi-attribute, decision-making approach based on the reasoning, knowledge, and experience of the experts in the field. ANP can act as a valuable aid for decision making involving both tangible as well as intangible attributes that are associated with the model under study. ANP relies on the process of eliciting managerial inputs, thus allowing for a structured communication among decision makers. Thus, it can act as a qualitative tool for strategic decision-making problems [Ravia et al., 2005]. AHP method does not measure the possible dependencies among factors. It assumes that the factors presented in the hierarchical structure are independent; however, this is not always a reasonable assumption. The possible dependency among factors can only be determined as a result of internal and external environmental analyses. The ANP is a generalization of the AHP. While the AHP represents a framework with a uni-directional hierarchical AHP relationship, the ANP allows for complex interrelationships among decision levels and attributes [Yuksel et al., 2007]. For instance, not only does the importance of the criteria determine the importance of the alternatives, as in a hierarchy, but the importance of the alternatives may also have an impact on the importance of the criteria. Therefore, a hierarchical representation with a linear top-to-bottom structure is not suitable for a complex system. A system with feedback can be represented by a network. The structural differences between a hierarchy and a network are shown in Fig. 1.

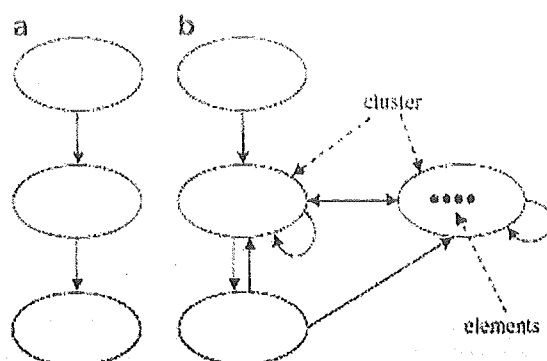


Fig. 1 Structural difference between a hierarchy and a network:
(a) a hierarchy; (b) a network.

3.1 Why ANP?

Although the AHP technique removes the deficiencies inherent in the measurement and evaluation steps of the problem analysis, it does not measure the possible dependencies among factors. The AHP method assumes that the factors presented in the hierarchical structure are independent; however, this is not always a reasonable presumption. The possible dependency among factors can only be determined as a result of internal and external environmental analyses.

3.2 Advantages of ANP [Yazgan, 2010]

- ANP is a comprehensive technique that allows for the inclusion of all the relevant criteria: tangible as well as intangible, which have some bearing on decision making process.
- ANP allows for more complex relationship among the decision levels and attributes as it does not require a strict hierarchical structure.
- In decision-making problems, it is very important to consider the interdependent relationship among the criteria because of the characteristics of interdependence that exist in real-life problems. The ANP methodology allows for the consideration of interdependencies among levels of criteria. Thus, it is a more desirable multi criteria decision-making tool.
- ANP is suitable in considering both qualitative as well as quantitative characteristics which need to be considered as well as taking nonlinear interdependent relationship among the attributes into consideration.
- ANP is unique in the sense that it provides synthetic scores, which is an indicator of the relative ranking of different alternatives available to the decision maker.
- Because of the increased number of pair wise comparison, the calculation process for ANP is more time consuming. Nevertheless, ANP is closer to real situation which has considered the feedback and interdependence among criteria. Therefore, it gives more flexibility for constructing a decision model; ANP is considered to be a promising method.

3.3 Disadvantages of ANP [Ravia et al., 2005]

- Identifying the relevant attributes of the problem and determining their relative importance in decision-making process requires extensive discussion and brainstorming sessions. Also, data acquisition is a very time intensive process for ANP methodology.
- ANP requires more calculations and formation of additional pair-wise comparison matrices as compared to the AHP process. Thus, a careful track of matrices and pair-wise comparisons of attributes is necessary.
- The pair-wise comparison of attributes under consideration can only be subjectively performed, and hence their accuracy of the results depends on the user's expertise knowledge in the area concerned.

4. Fuzzy ANP

The inability of ANP to deal with the impression in the pair wise comparison process has been improved in fuzzy ANP. Instead of a crisp value, fuzzy ANP applies a range of values to incorporate the decision maker's uncertainty. It enhances the potential of the ANP for dealing with imprecise and uncertain human comparison judgments. [Ramik, 2007] developed a decision system using ANP and fuzzy inputs. In this paper extended arithmetic operations with fuzzy numbers are proposed as well as ordering fuzzy relations to compare fuzzy outcomes. [kaur et al., 2008] developed a fuzzy ANP-based approach for selecting ERP vendors. In this paper ANP equipped with fuzzy logic helps in overcoming the impreciseness and vagueness in the performance. [Wu et al., 2008] developed a fuzzy ANP-based approach to evaluate medical organizational performance. This paper proposes an evaluation model using fuzzy analytic network process (FANP). The proposed model can provide Taiwan's hospital accreditation policy a reference material, making it highly applicable for academic and government purposes. [Rafiei et al., 2009] developed an ordered partitioning in hybrid MTS/MTO contexts using fuzzy ANP. In this paper, a model based on analytic network process is developed to tackle the addressed decision. Since the regarded decision deals with the uncertainty and ambiguity of data as well as experts' and managers' linguistic judgments, the proposed model is equipped with fuzzy sets theory. [Rouyendegh et al., 2010] developed the DEA - fuzzy ANP department ranking model applied in Iran Amirkabir University. This research is a two-stage model designed to fully rank the organizational departments where each department has multiple inputs and outputs.

5. A Gaussian Fuzzy ANP Proposed Model (GFANP)

5.1 Problem Formulation

It is required to develop a decision-making system, which helps decision-makers in the Egyptian government to select the best alternatives for the different scenarios of electrical power generation in Egypt. The highest priority would be the best (see fig. 2). There are 3 alternative scenarios:

Alt#1: the current one,

Alt#2: 20% nuclear, 75% petrol, 5% other, and

Alt#3: 25% nuclear, 65% petrol, 5% solar, 5% other.

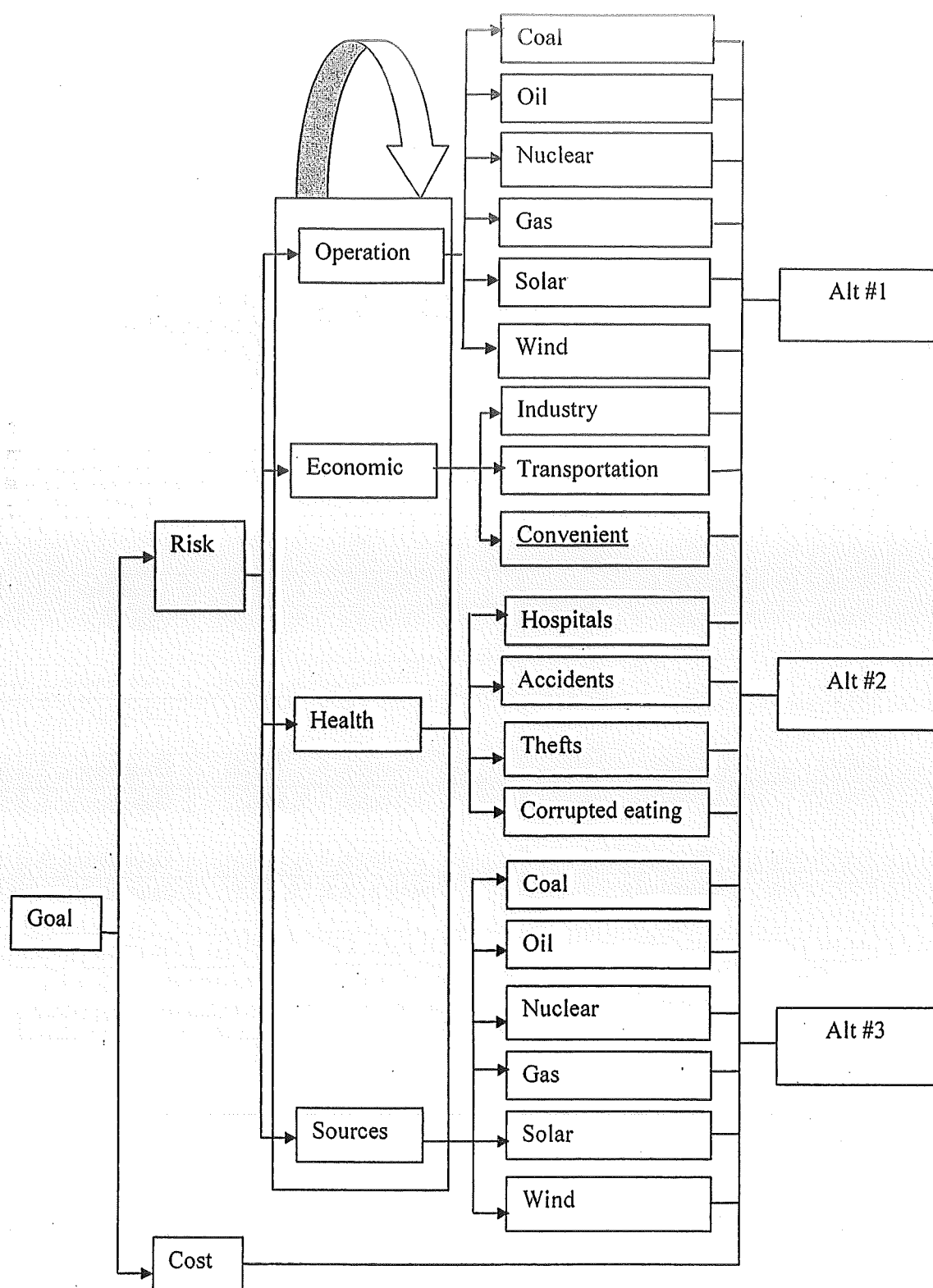


Fig. 2 The ANP Model

5.2 The Proposed Model

The proposed model to overcome the problem of triangular fuzzy numbers depends on replacing them by Gaussian fuzzy numbers. It is clear that defining Gaussian fuzzy numbers over the preference scale results in real intersection between any number and all the other numbers. This eliminates the problem of getting some alternatives have the same rank and consequently be treated equivalently. Thus, after illustrating that idea, we propose a modified preference table, shown in table 1, in which we introduce Gaussian fuzzy numbers to the eleven-point scale table. It should be noted that the centers (μ 's) of the Gaussian preference values must be the same as the crisp preference scale values. However, the widths (σ 's) can be freely assumed according to the existing amount of uncertainty.

Table 1: The preference table: μ = crisp number, $\sigma = 0.5$

LINGUISTIC VARIABLE	Crisp No.	Triang(x, a, b, c)	Gaussian(x, μ , σ)
Equally preferred (EP)	1	Triang(x, 1, 1, 1)	Gaussian(x, 1, 0.5)
Equally to Weakly preferred (EWP)	2	Triang(x, 1.5, 2, 2.5)	Gaussian(x, 2, 0.5)
Weakly preferred (WP)	3	Triang(x, 2.5, 3, 3.5)	Gaussian(x, 3, 0.5)
Weakly to Moderately preferred (WMP)	4	Triang(x, 3.5, 4, 4.5)	Gaussian(x, 4, 0.5)
Moderately preferred (MP)	5	Triang(x, 4.5, 5, 5.5)	Gaussian(x, 5, 0.5)
Moderately to strongly preferred (MSP)	6	Triang(x, 5.5, 6, 6.5)	Gaussian(x, 6, 0.5)
Strongly preferred (SP)	7	Triang(x, 6.5, 7, 7.5)	Gaussian(x, 7, 0.5)
Strongly to very strongly preferred (SVP)	8	Triang(x, 7.5, 8, 8.5)	Gaussian(x, 8, 0.5)
Very strongly preferred (VP)	9	Triang(x, 8.5, 9, 9.5)	Gaussian(x, 9, 0.5)
Very strongly to extremely preferred (VEP)	10	Triang(x, 9.5, 10, 10.5)	Gaussian(x, 10, 0.5)
Extremely preferred (XP)	11	Triang(x, 10.5, 11, 11.5)	Gaussian(x, 11, 0.5)

Gaussian functions have the advantage of being fully determined using only two parameters, i.e. center (μ) and width (σ) and its value never equals to zero (within the range of the preference scale values). Thus, the intersection must be existed between every fuzzy number and all the others. In this case, shortages of the triangular fuzzy number are overcome.

The definition of Gaussian function is as follows:

$$\text{Gaussian}(x; \mu, \sigma) = \exp \left[\frac{-(x - \mu)^2}{\sigma^2} \right] \quad (1)$$

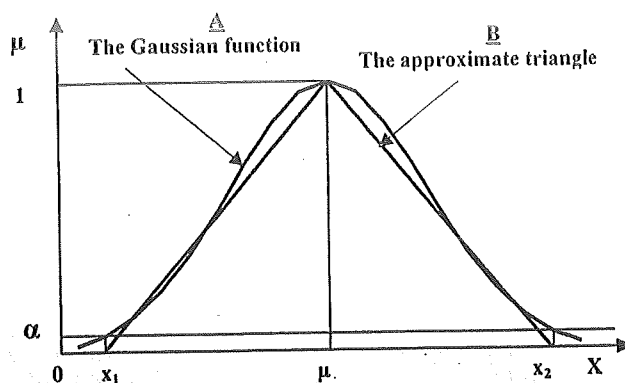


Fig. 3 Gaussian Function A and its approximated triangle B

At any level α , as in Fig. 3, it is shown that:

$$\alpha = \exp \left[\frac{-(x-\mu)^2}{\sigma^2} \right] \quad (2)$$

$$\therefore x_1 = \mu - \sigma \times \sqrt{-\ln(\alpha)} \quad \text{and} \quad (3)$$

$$x_2 = \mu + \sigma \times \sqrt{-\ln(\alpha)} \quad (4)$$

It is clear that as long as α level is small enough, then it is possible to get a good fuzzy triangular approximation of $G(x, \mu, \sigma)$ by $T(x, x_1, \mu, x_2)$. Such an approximation is useful for performing the fuzzy arithmetic operations to get S_i as shown in equation (5). Once, we get S_i 's as triangle fuzzy numbers, they can be returned back to Gaussian to perform the ranking step.

For example suppose that

$\sigma = 0.5$ and $\alpha = 0.1$ then

$x_1 = \mu - 0.76$ and

$x_2 = \mu + 0.76$

The steps of the modified fuzzy ANP (FANP) method are illustrated as follows:

Let G_{ij} be the elements of the preference matrix after performing the triangular approximation, then:

Step 1:

$$S_i = \frac{\sum_j G_{ij}}{\sum_i \sum_j G_{ij}} \quad (5)$$

$$= \frac{\sum_j (l_i^j, m_i^j, u_i^j)}{\sum_i \sum_j (l_i^j, m_i^j, u_i^j)} \quad (6)$$

Where

$$l_i^j \cong m_i^j - \sigma_i^j \sqrt{-\ln(\alpha)} \quad (7)$$

$$u_i^j \cong m_i^j + \sigma_i^j \sqrt{-\ln(\alpha)} \quad (8)$$

To get good triangular approximation, we choose a low level for α .

For example let $\alpha = 0.001$

$$\therefore S_i = \frac{\left(\sum_j l_i^j, \sum_j m_i^j, \sum_j u_i^j \right)}{\left(\sum_i \sum_j l_i^j, \sum_i \sum_j m_i^j, \sum_i \sum_j u_i^j \right)} \quad (9)$$

$$= \left(\frac{\sum_j l_i^j}{\sum_i \sum_j u_i^j}, \frac{\sum_j m_i^j}{\sum_i \sum_j m_i^j}, \frac{\sum_j u_i^j}{\sum_i \sum_j l_i^j} \right) \quad (10)$$

$$\sum_j l_i^j = \sum_j m_i^j - \sum_j \sigma_i^j (\sqrt{-\ln(\alpha)}) \quad (11)$$

$$\sum_j u_i^j = \sum_j m_i^j + \sum_j \sigma_i^j (\sqrt{-\ln(\alpha)}) \quad (12)$$

$$\sum_i \sum_j l_i^j = \sum_i \sum_j m_i^j - \sum_i \sum_j \sigma_i^j (\sqrt{-\ln(\alpha)}) \quad (13)$$

$$\sum_i \sum_j u_i^j = \sum_i \sum_j m_i^j + \sum_i \sum_j \sigma_i^j (\sqrt{-\ln(\alpha)}) \quad (14)$$

$$\therefore S_i = (x_{s_i}^L, m_{s_i}, x_{s_i}^R)$$

Where

$$m_{s_i} = \frac{\sum_j m_i^j}{\sum_i \sum_j m_i^j} \quad (15)$$

$$x_{s_i}^L = \frac{\sum_j l_i^j}{\sum_i \sum_j u_i^j} \quad (16)$$

$$x_{s_i}^R = \frac{\sum_j u_i^j}{\sum_i \sum_j l_i^j} \quad (17)$$

Now, S_i can be returned back to a Gaussian fuzzy number (but asymmetric in this case) as follows:

$$\sigma_{s_i}^L = \frac{m_{s_i} - x_{s_i}^L}{\sqrt{-\ln(\alpha)}} \quad (18)$$

$$\sigma_{s_i}^R = \frac{x_{s_i}^R - m_{s_i}}{\sqrt{-\ln(\alpha)}} \quad (19)$$

Where $\sigma_{s_i}^L$: is the width of the left branch of the Gaussian fuzzy number.

$\sigma_{s_i}^R$: is the width of the right branch of the Gaussian fuzzy number.

Now, S_i becomes an asymmetric Gaussian number as follows:

$$\mu_{s_i}(x) = \begin{cases} \exp\left[-\left(\frac{x-m_{s_i}}{\sigma_{s_i}^L}\right)^2\right] & \text{if } x \leq m_{s_i} \\ \exp\left[-\left(\frac{x-m_{s_i}}{\sigma_{s_i}^R}\right)^2\right] & \text{if } x > m_{s_i} \end{cases} \quad (20)$$

Step 2:

Let $\mu_1(x)$ and $\mu_2(x)$ be two Gaussian fuzzy numbers having the following forms:

$$\mu_{s_1}(x) = \begin{cases} \exp\left[-\left(\frac{x-m_{s_1}}{\sigma_{s_1}^L}\right)^2\right] & \text{if } x \leq m_{s_1} \\ \exp\left[-\left(\frac{x-m_{s_1}}{\sigma_{s_1}^R}\right)^2\right] & \text{if } x > m_{s_1} \end{cases}, \quad (21)$$

and

$$\mu_{s_2}(x) = \begin{cases} \exp\left[-\left(\frac{x-m_{s_2}}{\sigma_{s_2}^L}\right)^2\right] & \text{if } x \leq m_{s_2} \\ \exp\left[-\left(\frac{x-m_{s_2}}{\sigma_{s_2}^R}\right)^2\right] & \text{if } x > m_{s_2} \end{cases} \quad (22)$$

The intersection point between two Gaussian functions is shown in Fig. 4.

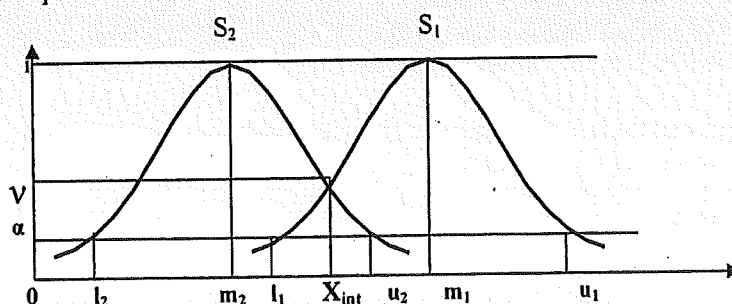


Fig. 4 Intersection Point Between Two Gaussian Functions

$$V = \begin{cases} \exp \left[- \left(\frac{(m_{s_2} - m_{s_1})^2}{(\sigma_{s_1}^L + \sigma_{s_2}^R)} \right) \right] & \text{if } m_{s_1} > m_{s_2} \\ \exp \left[- \left(\frac{(m_{s_2} - m_{s_1})^2}{(\sigma_{s_1}^R + \sigma_{s_2}^L)} \right) \right] & \text{if } m_{s_1} < m_{s_2} \end{cases} \quad [\text{Hefny, 2007}]. \quad (23)$$

The degree of possibility of $S_2 = \mu_{s_2}(x) \geq S_1 = \mu_{s_1}(x)$ is defined as

$$V(S_2 \geq S_1) = \text{hgt}(S_1 \cap S_2) = \mu_{s_2}(x_{\text{int}}) \quad (24)$$

$$V(S_2 \geq S_1) = \begin{cases} 1 & \text{if } m_{s_2} \geq m_{s_1}, \\ \exp \left[- \left(\frac{(m_{s_2} - m_{s_1})^2}{(\sigma_{s_2}^R + \sigma_{s_1}^L)} \right) \right] & \text{if } m_{s_2} < m_{s_1} \end{cases} \quad (25)$$

where x_{int} is the ordinate of the inner intersection point between $\mu_{s_2}(x)$ and $\mu_{s_1}(x)$. To compare S_1 and S_2 , the values of both $V(S_2 \geq S_1)$ and $V(S_1 \geq S_2)$ are needed.

Step 3:

The degree of possibility for a Gaussian fuzzy number S_i to be greater than k Gaussian fuzzy numbers S_i ($i = 1, 2, \dots, k$) can be defined by

$$\begin{aligned} V(S > S_1, S_2, \dots, S_k) \\ &= V[(S > S_1) \text{ and } (S > S_2) \text{ and } \dots \text{ and } (S > S_k)] \\ &= \min V(S > S_i), i = 1, 2, 3, \dots, k. \end{aligned} \quad (26)$$

Assume that

$$d'(A_i) = \min V(S_i > S_j) \text{ For } j = 1, 2, \dots, n; j \neq i. \quad (27)$$

Then the weight vector is given by:

$$W' = (d'(A_1), d'(A_2), \dots, d'(A_n))^T, \quad (28)$$

where A_i ($i = 1, 2, \dots, n$) are n elements.

Step 4:

Via normalization, the normalized weight vectors are

$$W = (d(A_1), d(A_2), \dots, d(A_n))^T, \quad (29)$$

$$\text{where } d(A_i) = \frac{d'(A_i)}{\sum_i d'(A_i)} \quad (30)$$

This gives the required priority weights of one alternative among others.

6. Experimental Results and Discussion

It is needed to rank the alternatives A_1, A_2, A_3, A_4, A_5 over Criteria C_1, C_2, C_3, C_4 . According to fig. 4, the linguistic preference matrices of the different criteria nodes are given in tables 2 to 7. On the other hand, the inner dependences matrices with respect to different criteria nodes are given in tables 8 to 11. The inner dependence among factors is shown in Fig. 5.

Table 2: The evaluation matrix with respect to the Goal

	Risk	Cost
Risk	EP	WMP
Cost		EP

Table 3: The evaluation matrix with respect to the Risk

	Operation	Economic	Health	Sources
Operation	EP	WMP	EWP	WP
Economic		EP	EWP	EWP
Health			EP	EP
Sources				EP

Table 4: The evaluation matrix with respect to the Operation Risk

	Coal	Oil	Nuclear	Gas	Solar	Wind
Coal	EP					
Oil	WMP	EP	EP			
Nuclear	EWP		EP			
Gas	WP	EWP	EWP	EP		
Solar	MP	WP	EWP	EWP	EP	
Wind	WMP	WP	EWP	EWP	EWP	EP

Table 5: The evaluation matrix with respect to the Economic Risk

	Industry	Transportation	Convenient
Industry	EP	EP	EP
Transportation		EP	EP
Convenient			EP

Table 6: The evaluation matrix with respect to the Health Risk

	Hospitals	Accidents	Thefts	Corr. Eating
Hospitals	EP	WP	EWP	EWP
Accidents		EP	WP	EWP
Thefts			EP	EP
Corr. Eating				EP

Table 7: The evaluation matrix with respect to the Sources Risk

	Coal	Oil	Nuclear	Gas	Solar	Wind
Coal	EP					
Oil	EWP	EP				
Nuclear	WP	EWP	EP			
Gas	WP	EWP	EWP	EP		
Solar	MP	WMP	WP	EWP	EP	EP
Wind	MP	WMP	WP	EWP		EP

$$w_1 = \begin{bmatrix} \text{Operation} \\ \text{Economic} \\ \text{Health} \\ \text{Sources} \end{bmatrix} = \begin{bmatrix} 0.30 \\ 0.27 \\ 0.23 \\ 0.20 \end{bmatrix}$$

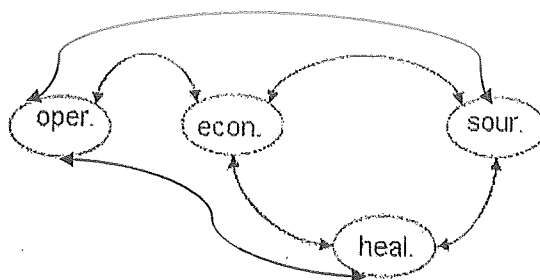


Fig. 5 Inner dependence among factors.

Inner relationship

Table 8: The inner dependence matrix with respect to the Operation

	Economic	Health	Sources
Economic	EP	EP	
Health		EP	
Sources	EWP	EWP	EP

Table 9: The inner dependence matrix with respect to the Economic

	Operation	Health	Sources
Operation	EP		EWP
Health	EWP	EP	EWP
Sources			EP

Table 10: The inner dependence matrix with respect to the Health

	Operation	Economic	Sources
Operation	EP		EWP
Economic	EWP	EP	EWP
Sources			EP

Table 11: The inner dependence matrix with respect to the Sources

	Operation	Economic	Health
Operation	EP		EWP
Economic	EWP	EP	EWP
Health			EP

$$W_2 = \begin{bmatrix} 1.00 & 0.33 & 0.33 & 0.33 \\ 0.25 & 1.00 & 0.48 & 0.48 \\ 0.25 & 0.48 & 1.00 & 0.19 \\ 0.50 & 0.19 & 0.19 & 1.00 \end{bmatrix} \quad (31)$$

$$V_{\text{factors}} = W_2 * w_1 = \begin{bmatrix} 1.00 & 0.33 & 0.33 & 0.33 \\ 0.25 & 1.00 & 0.48 & 0.48 \\ 0.25 & 0.48 & 1.00 & 0.19 \\ 0.50 & 0.19 & 0.19 & 1.00 \end{bmatrix} * \begin{bmatrix} 0.30 \\ 0.27 \\ 0.23 \\ 0.20 \end{bmatrix} = \begin{bmatrix} 0.27 \\ 0.28 \\ 0.23 \\ 0.22 \end{bmatrix} \quad (32)$$

Sample of the pair-wise comparisons between alternatives Alt.#1, Alt.#2, and Alt.#3 over criteria are given in tables from 12 to 18.

Alternatives

Table 12: The evaluation matrix with respect to the Cost

	Alt. #1	Alt. #2	Alt. #3
Alt. #1	EP	EWP	WP
Alt. #2		EP	
Alt. #3		EWP	EP

Table 13: The evaluation matrix with respect to the Coal

	Alt. #1	Alt. #2	Alt. #3
Alt. #1	EP		
Alt. #2	EWP	EP	
Alt. #3	WP	EWP	EP

Table 14: The evaluation matrix with respect to the Oil

	Alt. #1	Alt. #2	Alt. #3
Alt. #1	EP		
Alt. #2	EWP	EP	
Alt. #3	WP	WP	EP

Table 15: The evaluation matrix with respect to the Nuclear

	Alt. #1	Alt. #2	Alt. #3
Alt. #1	EP		
Alt. #2	EWP	EP	
Alt. #3	EWP	EWP	EP

Table 16: The evaluation matrix with respect to the Gas

	Alt. #1	Alt. #2	Alt. #3
Alt. #1	EP		
Alt. #2	WP	EP	
Alt. #3	WP	EWP	EP

Table 17: The evaluation matrix with respect to the Solar

	Alt. #1	Alt. #2	Alt. #3
Alt. #1	EP		
Alt. #2	WP	EP	
Alt. #3	EWP	EWP	EP

Table 18: The evaluation matrix with respect to the Wind

	Alt. #1	Alt. #2	Alt. #3
Alt. #1	EP		
Alt. #2	WP	EP	
Alt. #3	WP	WP	EP

Then, the overall normalized priority weight vector of the alternatives is obtained as follows:

$$W = [0.25, 0.33, 0.42]^T, \quad (33)$$

which means that Alt.#3 is the best alternative.

7. Conclusions

In the proposed model it is possible to benefit from the advantages of both interval and fixed value judgments. Shortages caused by each of them can be avoided. GFANP model provides expert judgments the flexibility of using interval values in their preference matrices instead of crisp values. Gaussian fuzzy numbers are used instead of triangular numbers. By using them the case of zero weights will never exist.

8. References

[Chen et al., 2001]

Chen, M. Y., and D. A. Linkens, "A Systematic Neuro-Fuzzy Modeling Framework With Application to Material Property Prediction", IEEE Transactions Systems Manufacturing, Cybernetics, Part B, vol. 31, no. 5, pp. 781-790, 2001.

[Ha et al., 2011]

D.W. Ha, J.M. Kwon, S.K. Baik, Y.J. Lee, K.S. Han, R.K. Ko, M.H. Sohn, K.C. Seong, "Purification of condenser water in thermal power station by superconducting magnetic separation", Physica C: Superconductivity, Vol. 471, Issues 21–22, pp. 1530-1532, 2011.

[ELECG, 2010]

"Electricity Generation", last visit on October 2012, available at: http://en.wikipedia.org/wiki/Electricity_generation

[Gao et al., 2011]

Hui Gao, Xiangdong Jiao, Chanfeng Zhou, Qiuping Shen, Yan Yu , "Study on Remote Control Underwater Welding Technology Applied in Nuclear Power Station", *Procedia Engineering*, Vol. 15, pp. 4988-4993, 2011.

[HDELT, 2011]

"Hydroelectricity", last visit on October 2012, available at:

<http://en.wikipedia.org/wiki/Hydroelectricity>

[Hefny, 2007]

Hefny, H. A., "Comments on "Distinguishability Quantification of Fuzzy Sets"", *Information Sciences*, vol. 177, pp. 4832 - 4839, 2007.

[Jiang et al., 2008]

Zhibing Jiang, Jiangning Zeng, Quanzhen Chen, Yijun Huang, Xiaoqun Xu, Yibo Liao, Lu Shou, Jingjing Liu, "Tolerance of copepods to short-term thermal stress caused by coastal power stations", *Journal of Thermal Biology*, Vol. 33, No. 7, pp. 419-423, 2008.

[kaur et al., 2008]

Prabjot kaur, N. C. Mahanti, "A Fuzzy ANP-Based Approach For Selecting ERP Vendors", *International Journal of Soft Computing*, vol. 3, no. 1, pp. 24-32, 2008.

[MBendi, 2010]

"Electrical Power in Egypt", last visit on October 2012, available at:

<http://www.mbendi.com/indy/powr/af/eg/p0005.htm>

[Popov, 2011]

Dimitry Popov , "An option for solar thermal repowering of fossil fuel fired power plants", *Solar Energy*, Vol. 85, No. 2, pp. 344-349, 2011.

[Rafiei et al., 2009]

H. Rafiei, M. Rabbani, "Order Partitioning in Hybrid MTS/MTO Contexts using Fuzzy ANP", *World Academy of Science, Engineering and Technology*, vol. 58, 2009.

[Ramik, 2007]

Jaroslav Ramik, "A Decision System Using ANP and Fuzzy Inputs", *International Journal of Innovative Computing, Information and Control*, vol. 3, no. 4, pp. 825-837, 2007.

[Ravia et al., 2005]

V. Ravia, Ravi Shankara, M. K. Tiwarib, "Analyzing alternatives in reverse logistics for end-of-life computers: ANP and balanced scorecard approach", *Computers & Industrial Engineering*, vol. 48, pp. 327-356, 2005.

[Rouyendegh et al., 2010]

Babak Daneshvar Rouyendegh, Serpil Erol, "The DEA - Fuzzy ANP Department Ranking Model Applied in Iran Amirkabir University", *Acta Polytechnica Hungarica*, vol. 7, no. 4, 2010.

[Sahu et al., 2009]

S.K. Sahu, R.C. Bhangare, P.Y. Ajmal, S. Sharma, G.G. Pandit, V.D. Puranik ,
"Characterization and quantification of persistent organic pollutants in fly ash from coal
fueled thermal power stations in India", Microchemical Journal, Vol. 92, No. 1, pp. 92-
96, 2009.

[Smoot et al., 2003]

L.Douglas Smoot, Larry L. Baxter, "Fossil Fuel Power Stations—Coal Utilization",
Encyclopedia of Physical Science and Technology (Third Edition), pp. 121-144, 2003.

[SOLEG, 2010]

"Solar Energy", last visit on October 2012, available at:
http://en.wikipedia.org/wiki/Solar_energy

[Sovacool et al., 2011]

Benjamin K. Sovacool, Saroj Dhakal, Olivia Gippner, Malavika Jain Bambawale,
"Halting hydro: A review of the socio-technical barriers to hydroelectric power plants in
Nepal", Energy, Vol. 36, No. 5, pp. 3468-3476, 2011.

[Visschers et al., 2011]

Vivianne H.M. Visschers, Carmen Keller, Michael Siegrist , "Climate change benefits
and energy supply benefits as determinants of acceptance of nuclear power stations:
Investigating an explanatory model", Energy Policy, Vol. 39, No. 6, pp. 3621-3629,
2011.

[WINEG, 2010]

"Wind Power", last visit on October 2012, available at:
http://en.wikipedia.org/wiki/Wind_power

[Wu et al., 2001]

Wu, S., M. J. Er, and Y. Gao, "A Fast Approach for Automatic Generation of Fuzzy
Rules by Generalized Dynamic Fuzzy Neural Networks", IEEE Transactions on Fuzzy
Systems, vol. 9, no. 4, pp. 578-594, 2001.

[Wu et al., 2008]

Cheng-Ru Wu, Che-Wei Chang, Hung-Lung Lin, "A Fuzzy ANP-based Approach to
Evaluate Medical Organizational Performance", Information and Management Sciences,
vol. 19, no. 1, pp. 53-74, 2008.

[Xu et al., 2011]

S.M. Xu, X.D. Huang, R. Du , "An investigation of the solar powered absorption
refrigeration system with advanced energy storage technology", Solar Energy, Vol. 85,
No. 9, pp. 1794-1804, 2011.

[Yazgan, 2010]

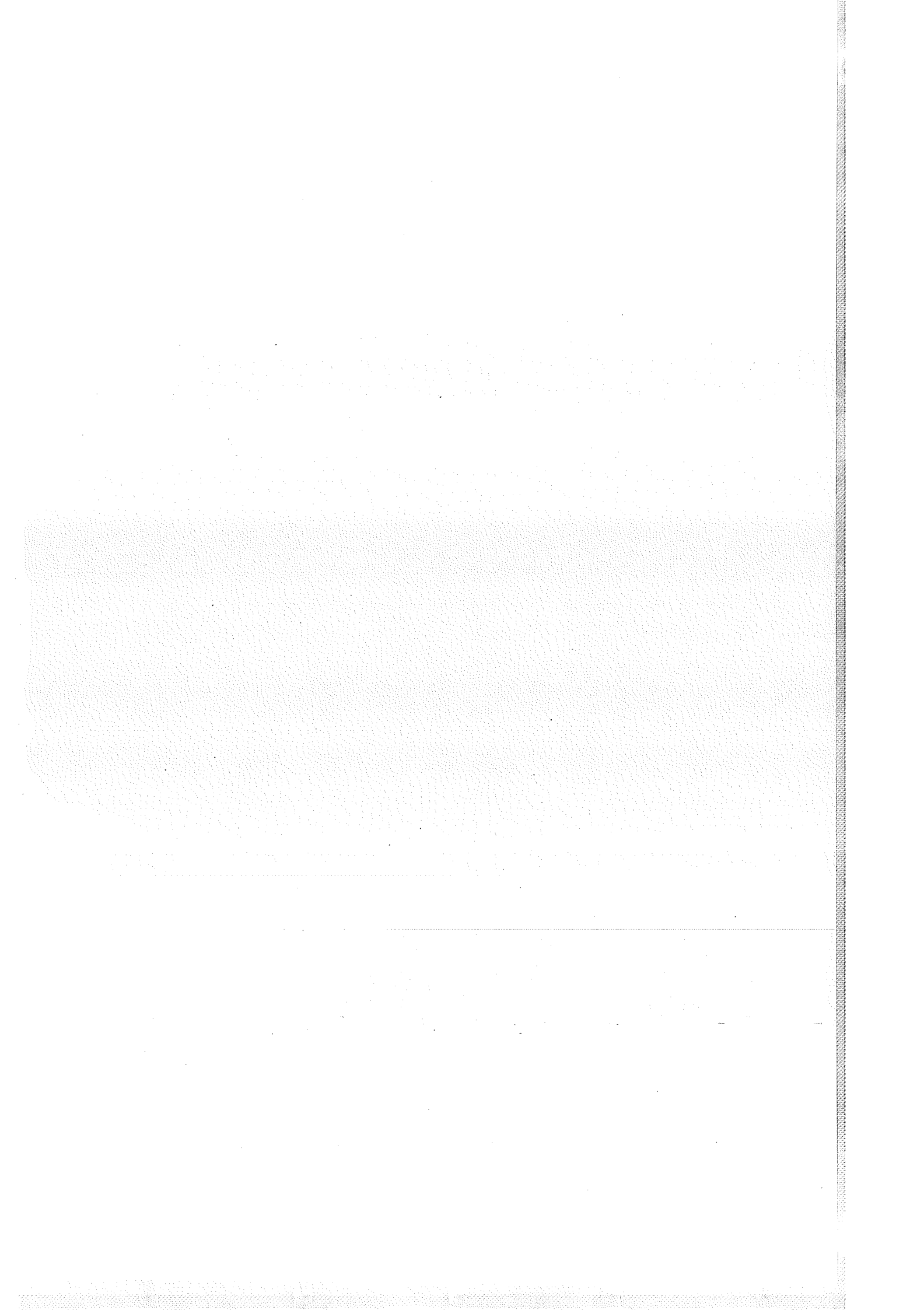
Harun Resit Yazgan, "Selection of dispatching rules with fuzzy ANP approach",
International Journal of Advanced Manufactured Technologies, 2010.

[Yuksel et al., 2007]

Ihsan Yuksel, Metin Dagdeviren, "Using The Analytic Network Process (ANP) in a SWOT Analysis – A Case Study For a Textile Firm", Information Sciences, vol. 177, pp. 3364–3382, 2007.

[Zhang et al., 2012]

Rui Zhang, Jianzhong Zhou, Yongqiang Wang, "Multi-objective optimization of hydrothermal energy system considering economic and environmental aspects Original", International Journal of Electrical Power & Energy Systems, Vol. 42, Issue 1, pp. 384-395, 2012.



Reverse Engineering of PHP Grammar

Mohamed M.S.¹ El-Licy F.A.¹ Salah A.I.²

Abstract:

This paper presents an extracted regular expressions and context free grammar for the PHP programming language. The grammar was established as a formal specification of the PHP grammar to fill up the inadequacies needed to build a PHP front end compiler. The grammar and the utilization of the generated parser could serve for the analysis, and evaluation of PHP code.

The extracted regular expressions of PHP programming Language are utilized to generate context free grammars.

A context free grammar is written to support predicative parsing. PHP code samples are parsed using the generated predicative parser to exercise the grammar.

The generated parse tree will be useful for verifying the security of code. Also, it could be used in creating a parse tree and unparsing it to another programming language (e.g. asp).

Keywords: Regular Expression, PHP programming language, predicative parsing, context free grammar, PHP grammar extraction, PHP Regular Expression Extraction.

1. Introduction:

PHP is a server-side scripting language designed, specifically, for building web application. Within an HTML page, one can embed PHP code that will be executed each time the page is visited/loaded. The PHP code is interpreted at the Web server to generate HTML or other Client side web language [1, 2].

The formal specifications of the syntax and semantics of PHP are very limited. De Vries and Gilbert [3], stated that "no official grammar for PHP exists". However, The PHP lexical analyzer was, originally, defined using a Flex³ description, with Bison specification for the parser [4].

¹ Department of Computer and Information Sciences, Institute of statistical Studies and research, Cairo University

² Faculty of Computer science and Information, Computer Science Department, Cairo University

³ Flex is a tool for generating scanners. A scanner, sometimes called a tokenizer, is a program which recognizes lexical patterns in text.

Neither of these specifications is adequate as a formal description for PHP programming language.

The problem was the difficulties involved in extracting the formal specification of PHP language from Bison's. That is because of the redundancy and the weak structuring of its grammar rules. Bison's parser includes appropriately rules that are maintained at certain points during parsing such as: the convoluted rules which attempt to enforce static checks and the embedded semantic actions that ensure associativity, [3].

The problem tackled in this article is to generate a formal grammar for PHP programming Language. The generated grammar is amenable to be utilized for analyzing, refactoring and optimizing PHP codes. It, also, facilitates the translation of PHP code to other web languages such as ASP and JSP.

A PHP parser for the generated grammar is built as a predictive parser. The generated parse tree could be utilized for static and dynamic analysis of PHP code and could also be used for code unparsing into another conventional programming language. This paper presents samples of the extracted regular expressions and its construction into context free grammars⁴.

The paper is structured as follows: section 2 shows the related work, section 3, presents the research approach, section 4, introduces the reverse engineering of the PHP grammar, section 5, presents the lexical analyzer, section 6 introduces the predictive parser and shows some PHP examples, including a comparison with Bison and section 7 presents the conclusion and future work.

2. Related work

Some of the research papers that are related to the presented topics are discussed in the following subsections.

2.1 Design and Implementation of a PHP Compiler Front-end

The technical report [5], described the design and implementation of a front-end for *phc*, the open source *PHP* compiler. This front-end provides an excellent basis for developing tools which process PHP source code, and consists of a well-defined Abstract Syntax Tree (AST) specification for the PHP language.

⁴ Full grammar of the language is available and will be published in a Master dissertation by the 1st Author

2.2 Analyzing Ambiguity of Context-Free Grammars

Basten, in [6], discussed the usability of ambiguity detection methods in context-free grammars. He showed that it is a recurring problem in language design and parser generation, as well as in applications where grammars are used as models of real-world physical structures.

The authors, in [7], observed that there was a simple linguistic characterization of the grammar ambiguity problem, and they showed how to exploit this by presenting an ambiguity analysis framework based on conservative language approximations. As a concrete example, the authors proposed a technique based on local regular approximations and grammar unfolding.

2.3 Verified Decision Procedures on Context-Free Grammars

The authors of this paper verified three decision procedures on context-free grammars utilized in a program analyzer for a server-side programming language. One of the procedures decides the inclusion between a context-free language and a regular language. The other two decided decision problems related to the well-formedness and validity of XML documents [8]. From their formalization, the authors generated executable code for a balancedness checking procedure and incorporated it into an existing program analyzer [9].

2.4 Derivatives of Regular Expressions

Antimirov, in his paper [10], proposed a characterization of the structure of derivatives and proved several properties of derivatives for an expression with distinct symbols. Based on his work, Chen presented a simpler proof of the fact that the equation automaton is a quotient of the position automaton [11].

3. Research Approach

The approach of this research was motivated by an original goal of testing the vulnerability of the PHP code. It follows that, the specification of the syntax and semantics of PHP were not adequate to achieve the original goal. Therefore, a plan was established to extract the PHP grammar, which necessitates the definition of the language's tokens. Regular expressions are, then, generated for the tokens to support a formal

lexical analyzer. PHP context free grammar is, then, extracted using the PHP Programming language constructs. Figure 1, shows a block diagram for the process of extracting and generating the context free grammar.

The extracted grammar is written to support predicative parsing, which was selected because it is the most common technique for hand-generated parsers.

4. Grammar Reverse Engineering

Because of the lack of complete formal specifications for the PHP grammar a formal specifications was extracted, manually, from the language constructs. The objective was to gather the language constructs from a collection of PHP code in different web sites, books and PHP programming Manual. This collection was utilized, not only, to cover PHP constructs, but also to test and exercise the generated grammar.

Thereafter, a manual reverse engineering was performed to extract the PHP regular expression. The extracted regular expressions were applied to construct the PHP Context Free Grammar.

Figure 1, illustrates the steps involved in the extraction of the PHP context free grammar.

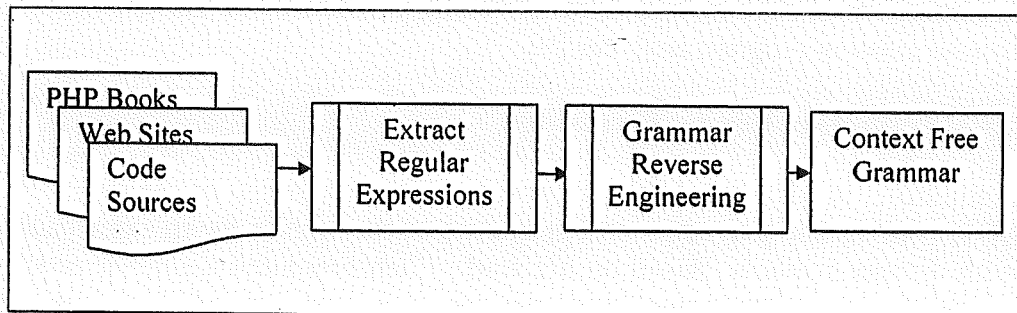


Figure 1: Grammar extraction

5. Lexical Analyzer

The lexical analyzer accepts PHP source code as input, and produces a stream of tokens as output, [12]. It was built by utilizing the list of regular expressions⁵ that have been extracted from the language.

⁵ Full list of regular expression of the language is available and will be published in a Master dissertation by the 1st Author.

5.1 The Regular Expressions

Each of the extracted constructs (token) was translated into a formal description in the form of a regular expression. *Microsoft .NET Framework* notation [13], was adopted to support the building of the regular expression. It was chosen because:

- It is one of the most popular regular expressions notation
- It is easy to be implemented using any of Microsoft .net programming languages.
- It incorporates the most popular features of other regular expression implementations such as those in Perl.
- It includes features that are not yet seen in other implementations, such as right-to-left matching, [13].

5.2 Examples of Regular Expressions

The following examples illustrate the generation of some of regular expressions for the extracted PHP tokens.

5.2.1 The variable name definition

A variable in PHP is represented by a dollar sign followed by the name of the variable. A valid variable name starts with a letter or underscore, followed by any number of letters, numbers, or underscores. [15]. Figure 2, shows the syntax diagram representing the identifier (variable) in PHP.

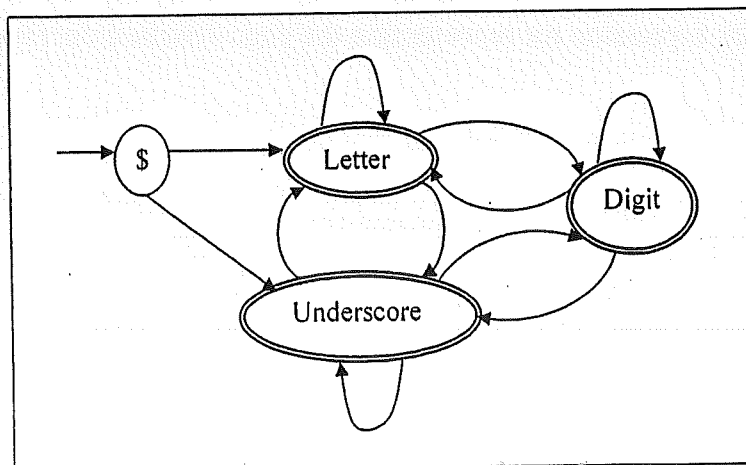


Figure 2: Identifier syntax diagram

The corresponding regular expression for the variable definition is:

`\$[a-zA-Z_][a-zA-Z0-9_]*` {reg1}

Note that ‘\’ is the escape character in the Microsoft .NET framework notation that was used because the ‘\$’ character has specific meaning in the description of the .NET language.

5.2.2 Concatenation operator

Concatenation operator is used in PHP to combine two strings of characters, [17]. Its regular expression is represented with the syntax diagram shown in Figure 3.

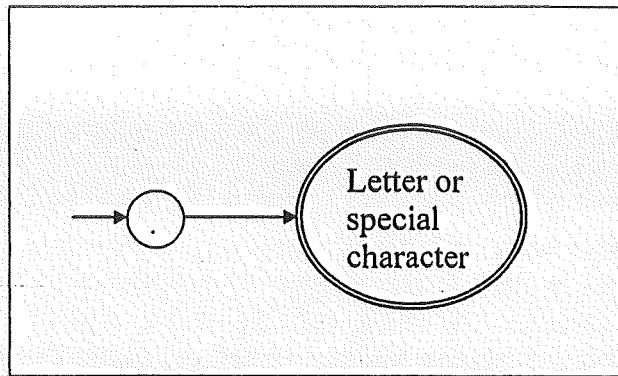


Figure 3: Concatenation operator syntax diagram

The regular expression for representing the concatenation operator is

`\.[^0-9]` {reg2}

This regular expression resolves two problematic cases:

1. The Dot character has its own meaning in the description language itself, therefore the usage of the escape character ‘\’.
2. Numerals digits following the dot always represent a decimal point, therefore the expression inside the square brackets [^0-9] are essential to avoid the occurrence of numeric digits after the Dot concatenation operator.

5.2.3 Numeric value

Numeric value represents constant number. The numbers in PHP can be written in one of the following ways: 132, 1.3, 0.768, .2 or 1.

The regular expression for the numeric value is

$[0-9]^* \backslash . ? [0-9]^*$

{reg 3}

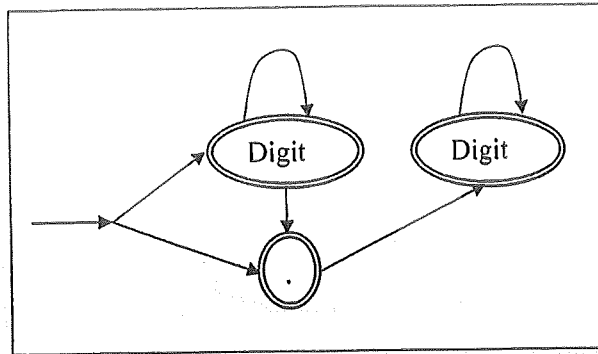


Figure 4: Numeric value syntax diagram

5.3 Ambiguity Solution

One string may match many regular expressions, which cause ambiguity problem for the lexical analyzer [14]. For example, the regular expression {reg 3}, specifies the numeric values '.2', '2.' and a single dot character '.' as numeric tokens. Also, the regular expression {reg 2}, specifies the dot character '.' as concatenation operator. In order for the lexical analyzer to distinguish between those two tokens, the regular expressions must be listed in certain order to overcome this and other ambiguity cases. That is to say, the lexical analyzer would check the concatenation operator regular expression before checking the numeric value regular expression, [12].

5.4 One Look-Ahead

For the lexical analyzer to recognize the correct token, it might need extra information about the next token in the string. This information is obtained by applying the one-look-ahead action. For example, to detect the built in function 'sqrt', the lexical analyzer must check that a left bracket is exist after the function name.

6. The Predictive Parser

The context free grammar was reverse engineered from the PHP language constructs. Its terminals are these tokens that were defined by regular expressions. The engineered grammar is then transformed into predicative grammar using techniques to eliminate left recursion and

perform left factoring, [12, 16]. Accordingly, a top-down predictive parser has been designed and implemented to execute the predictive grammar.

6.1 PHP Parsing Process

Predictive parser uses a stack to store the production rules associate with the current token. This rule will serve to define the appropriate production rule for the next token. Given a new token, the parser will check the parsing table for the appropriate production rule according to the non-terminal (*nt*) exist on the stack top. The stack will then be updated with the new production rules after popping the non-terminal (*nt*) from the stack.

The parsing table was, manually, created from the context free grammar. It stores the actions that the parser should take based on the input token and the value in the stack top. A parser function '*get next token*' prepares the input token, while the function '*Get Grammar rule*' takes the top non-terminal in the stack and the input token to fetch the parsing table for the next grammar rule. The function '*Update Stack*' accepts this grammar rule to update the stack.

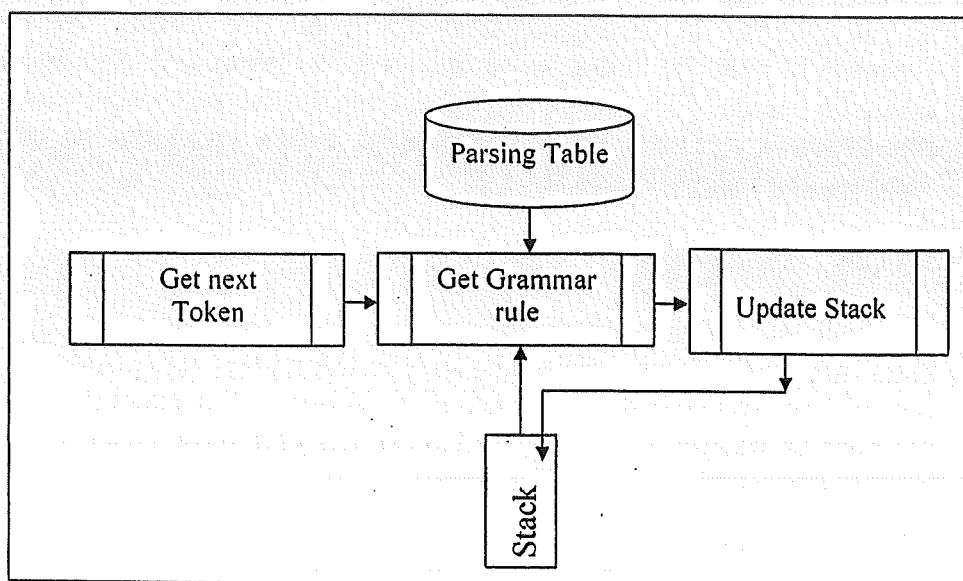


Figure 5: Predictive parser process

6.2 Abstract Syntax Trees for Some PHP Code Examples

The following PHP code writes the numbers from 1 to 5 using while loop statement.

{Code 1}

```
$i=1;
while($i<=5)
{
    echo "The number is " . $i . "<br />";
    $i+=1;
}
```

This code contains 3 types of statements

Assignment Statement

While Statement

Function Call (echo function)

The production rules for these statements are shown in the listing of {Grammar 1}. The terminals in {Grammar 1}, represent the tokens of the PHP programming language. The parser will apply this set of production rules to parse {Code 1} and generate the corresponding parse tree. The abstraction of the generated parse tree is shown in Figure 6.

Another example is shown in {Code 2}, which accepts a value from the user and writes either "Greater than five" or "Less than or equal five" accordingly.

{Code 2}

```
$x = $_GET["userinput"];
if($x>5)
    echo "Greater than 5";
else
    echo "Less than or equal 5";
```

{Grammar 1}

```

PROGRAM → STMTS
STMTS → STMT STMTS
        | ε
STMT → ASSIGNMENT ;
        | FUNCTION_CALL;
        | WHILE_STMT
ASSIGNMENT → IDENTIFIER ASSIGNMENT_FACT
ASSIGNMENT_FACT → = EXP
                | +=EXP
EXP → T EXP_LRE
EXP_LRE → ε
T → F T_LRE
T_LRE → . F T_LRE
        | ε
F → number
    | string
FUNCTION_CALL → echo ECHO_ARGS
IDENTIFIER → id
ECHO_ARGS → EXP ECHO_ARGS_FACT
ECHO_ARGS_FACT → ,EXP ECHO_ARGS_FACT
                | ε
LOGICAL_EXPR → LOGICAL_EXPR_T
LOGICAL_EXPR_LRE → ε
LOGICAL_EXPR_LRE → LOGICAL_EXPR_F
LOGICAL_EXPR_T_LRE → OP LOGICAL_EXPR_F
LOGICAL_EXPR_T_LRE → ε
LOGICAL_EXPR_F → EXP
                | LOGICAL_VALUE
                | (LOGICAL_EXPR)
OP → <=
BODY_STMTS → STMT
            | { STMTS }
WHILE_STMT → while(LOGICAL_EXPR) BODY_STMTS
    
```

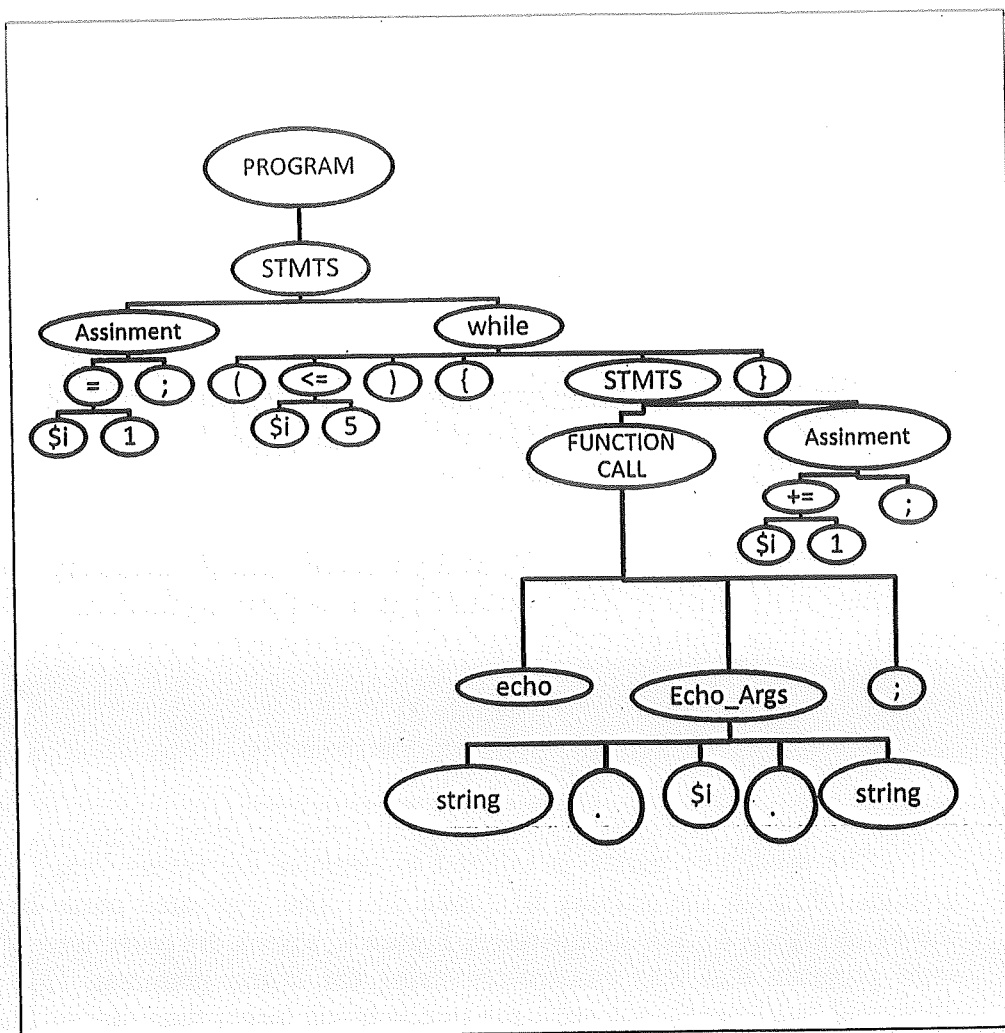


Figure 6: Abstract syntax tree for {Code 1}.

{Code 2} contains 3 types of statements:

Assignment Statement

If else Statement

Function Call (echo function)

{Grammar2} illustrates the production rules to be added to {Grammar 1} to parse {Code 2}. The Abstract parsing tree for code 2, is shown in Figure 7.

{Grammar 2}

```

STMT → ASSIGNMENT ;
      | FUNCTION_CALL;
      | IFSTMT
      | WHILE_STMT
OP → <=
    | >
IFSTMT → if(LOGICAL_EXPR) BODY_STMTS ELSE_PART
ELSE_PART → else BODY_STMTS
           | elseif (LOGICAL_EXPR) BODY_STMTS ELSE_PART
           | ε
    
```

6.3 Comparison with Bison PHP Grammar

The PHP code shown in {code 3}, contains a single statement which evaluates an expression. The expression is a method invocation; with the name '*printf*' that accepts the string "Hi" as parameter.

{Code 3}

```
printf("Hi");
```

The predictive parser is utilized to exercise the reverse engineered grammar through the parsing of {code 3}. The generated parse tree is shown in Figure 8. For comparison, Figure 9, shows the parse tree of the same program that was constructed using the Bison grammar, [3]. As illustrated in Figure 9, not only, the tree is much larger than that of Figure 8, but it, also, contains, many nodes that do not correspond to any conceptual notions. One will notice the repeated semicolons, as well as, the '*unticked_statement*' and '*r_variable*' notations for the code that does not contain any variables.

7. Conclusion and Future Work

The objective of the paper was fulfilled to extract a formal grammar for the PHP web programming Language. This grammar will fill up the formal inadequacies needed to build a PHP front end compiler. The grammar and the utilization of the generated predictive parser could serve for the analysis, and evaluation of PHP code.

This paper presents regular expressions and predictive grammar for PHP programming language. The generated regular expressions support the lexical analyzer to prevent the conflict ambiguity. A manually, generated

context free grammar was written and transformed to support predictive parsing. A Set of examples were presented to illustrate and compare the grammar and the generated parse trees to that of Bison's. The parse tree of the established grammar is in, general, much smaller and more conceptual than the Bison parse tree.

The regular expressions of PHP programming language were manually built from the language constructs. A reverse engineering of PHP context free grammar was established and transformed into a predictive grammar. A lexical analyzer and a predictive parser were designed and implemented using C# programming language. The parser was tested on a personal computer using Intel Pentium 4 - 3GH processor with 2GB of RAM.

The results established in this paper could, in particular, be applied to analyze and examine the PHP code-based web applications security. The parsed code of PHP programming language could be manipulated through testing techniques to detect web security vulnerabilities. These diagnoses should assist in identifying, not only the treatment for these security gaps, but also, the rules and conventions to seal/secure them.

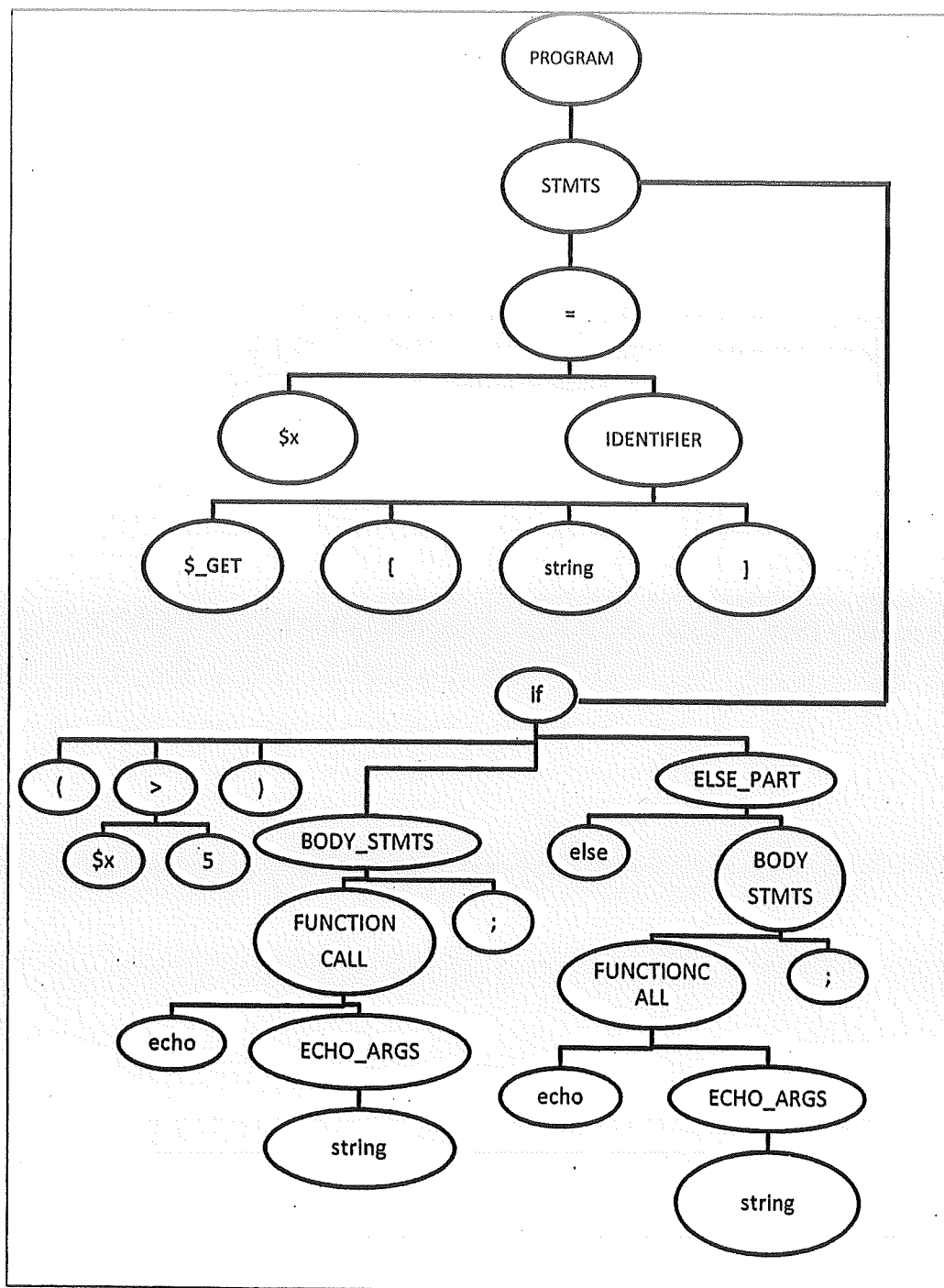


Figure 7: Abstract syntax tree for {Code 2}.

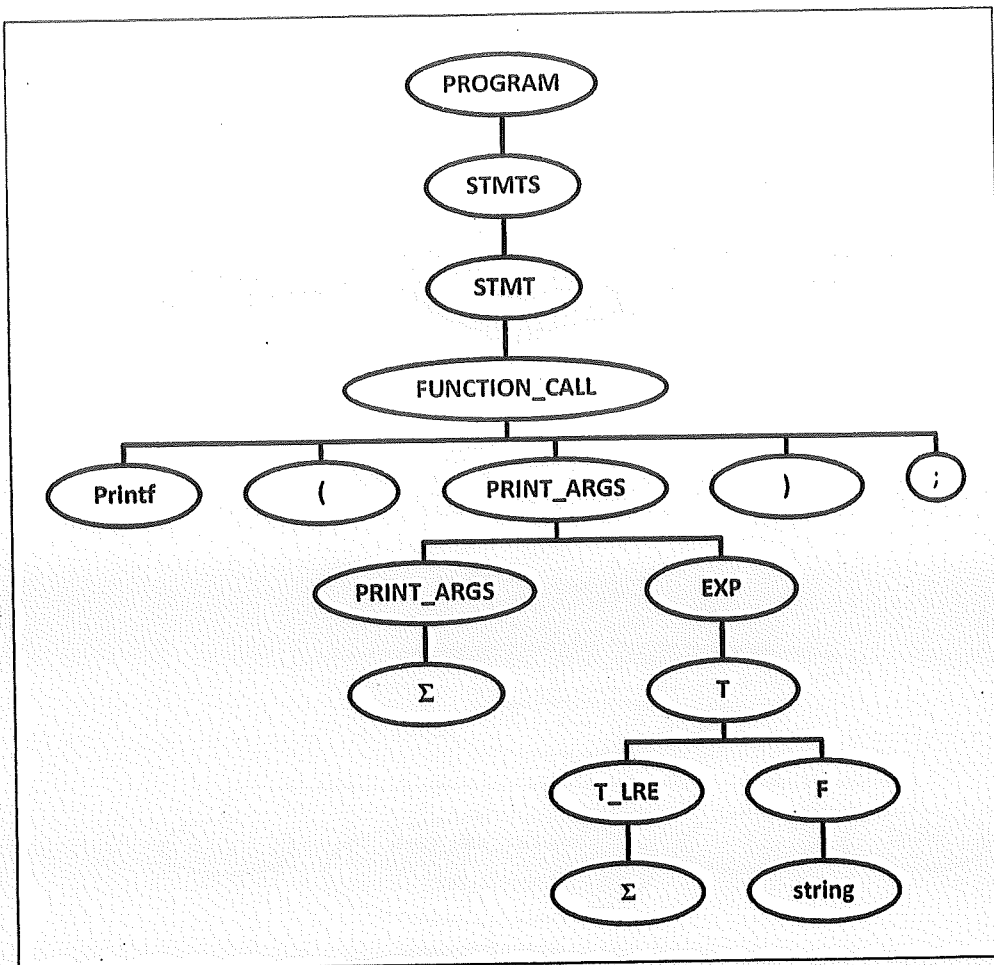


Figure 8: Parse tree of {Code 3} generated by the predictive parser

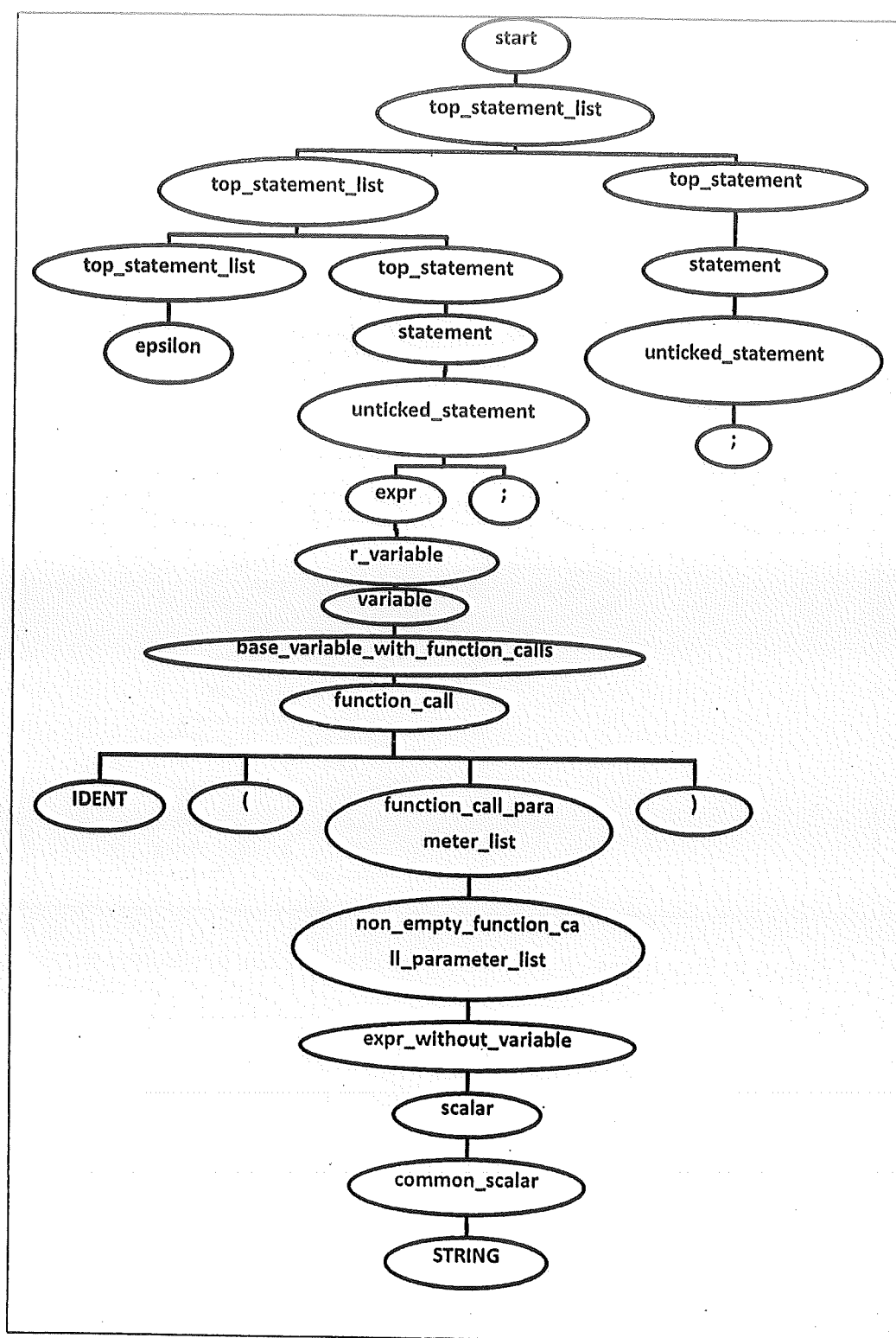


Figure 9: Parse tree of {Code 3} generated by Bison [3].

References

- [1] Janet Valade "PHP and MySQL For Dummies, 4th Edition", Dummies, USA, 2009.
- [2] Luke Welling and Laura Thomson "*PHP and MySQL Web Development* ", fourth Edition, Pearson Education, USA, 2009.
- [3] Edsko de Vries and John Gilbert "Design and Implementation of a PHP Compiler Front-end", Trinity College Dublin, September 21, 2007.
- [4] Free Software Foundation. Bison. Available at "<http://www.gnu.org/software/bison/manual/>", Retrieved on Nov 17.
- [5] The Open-source PHP Compiler "<http://www.phpcompiler.org/>", Retrieved on Nov 17, 2012.
- [6] Bas Basten "The Usability of Ambiguity Detection Methods for Context Free Grammars", *Electronic Notes in Theoretical Computer Science*, 2009.
- [7] Claus Brabrand, Robert Giegerich and Anders Moller "*Analyzing Ambiguity of Context-Free Grammars*", Springer-Verlag, 2010.
- [8] Jean Berstel and Luc Boasson "Formal Properties of XML Grammars and Languages", *Acta Informatica*. 2002.
- [9] Yasuhiko Minamide "Verified Decision Procedures on Context-Free Grammars", *the 20th International Conference on Theorem Proving in Higher Order Logics*, 2007.
- [10] Valentin Antimirov "Partial Derivatives of Regular Expressions and Finite Automaton Constructions", *Theoretical Computer Science*, 1996.
- [11] Haiming Chen and Sheng Yu "Derivatives of Regular Expressions and an Application," *Theoretical Computer Science and Applications Conference (WTCS 2012)*, LNCS 7160, 343-356, 2012.
- [12] Alfred Aho, Monica Lam, Ravi Sethi and Jeffrey Ullman "*Compilers - Principles, Techniques and Tools* ", Second Edition, Pearson Education, USA, June, 2006.

[13] .NET Framework Regular Expressions. Available at "[http://msdn.microsoft.com/en-us/library/hs600312\(v=vs.71\).aspx](http://msdn.microsoft.com/en-us/library/hs600312(v=vs.71).aspx)" Retrieved on Nov 17.

[14] Torben Aegidius Mogensen "*Basics of Compiler Design*". lulu, Denmark, 2010.

[15] PHP manual. Available at "<http://www.PHP.net/manual/en/>", Retrieved on Nov 17, 2012.

[16] Andrew Wilson Appel, "Modern Compiler Implementation in ML", Cambridge University Press, USA, 1997.

[17] Luke Welling and Laura Thomson, "*PHP and MySQL Web Development*", Fourth Edition, Pearson Education, USA, 2009.

Hand Gesture Recognition-Based Generative Versus Discriminative Models

Elmezain.M, El-Shinawy.S

Abstract

Many approaches to pattern recognition are founded on probability theory, and can be broadly characterized as either generative or discriminative according to whether or not the distribution of the image features. Generative and discriminative models have very different characteristics, as well as complementary strengths and weaknesses. In this paper, we study these models to recognize the patterns of alphabet characters (A-Z) and numbers (0-9). To handle isolated pattern, generative model as Hidden Markov Model (HMM) and Discriminative models like Conditional Random Field (CRF), Hidden Conditional Random Field (HCRF) and Latent-Dynamic Conditional Random Field (LDCRF) with different number of window size are applied on extracted pattern features. The gesture recognition rate is improved initially as the window size increase, but degrades as window size increase further. Experimental results show that the LDCRF is the best in terms of results than CRF, HCRF and HMM at window size equal 4. Additionally, our results show that; an overall recognition rates are 91.52%, 95.28%, 96.94% and 98.05% for CRF, HCRF, HMM and LDCRF respectively.

Keywords-Statistical Pattern Recognition; Generative Model, Discriminative Model, Human Computer Interaction.

1. Introduction

The hand gesture recognition is an active area of research in the vision community, mainly for the purpose of sign language recognition. Sign language recognition is an application area for Human Computer Interaction (HCI) to communicate with computers. The goal of pattern interpretation is to push the advanced human-computer communication to bring the performance of HCI close to human-human interaction. In the last decade, several methods of potential applications [1, 2, 3, 4, 5, 6] in the advanced hand gesture interfaces have been suggested but these differ from one another in their models. Some of these models are Neural Network [7], Hidden Markov Model (HMM) [1, 8, 9], Dynamic Time Warping (DTW) [10] and Conditional Random Field [3, 4, 5].

Elmezain et al. [1] proposed a real-time system to recognize American Sign Language (ASL) and numbers (0-9) using HMM. The experiment was performed with isolated hand motion trajectory that was employed for HMM for recognition and achieved 94.72% accuracy. *Yang et al.* [2] introduced an ASL recognition system based on a time-delay neural network. This system used the motion information to extract hand position where the recognition rate was 96.2%. *Yang et al.* [3] introduced a method for designing threshold model in a CRF model, which performs an adaptive threshold to distinguish between signs and non-sign patterns.

Department of Mathematics, Statistics & Computer Science Division, Faculty of Science, Tanta University, Egypt.

The experiments were performed with isolated and continuous data set according to extracted six features. *Sminchisescu et al.* [4] applied CRF model to recognize human motion activities and showed improvement over an HMM technique.

The difference between HMM and CRF is that HMM is generative model that defines a joint probability distribution to solve a conditional problem thus focusing on modeling the observation to compute the conditional probability $p(y|x)$. Moreover, one HMM is constructed per label (i.e. pattern) where HMM assumes that all the observation are independent. CRF uses an undirected graphical model to overcome the weakness of Maximum Entropy Markov Model (MEMM) [5]. CRF uses a single model of the joint probability of the labels sequence given the observation sequence. Therefore, there is trade-off in the weights of occurrences number of a feature value for each state [3]. Hidden Conditional Random Field (HCRF) is the extension of CRF that include hidden variables [11, 12]. HCRF can automatically model the local interconnection between labels (i.e. states) with hidden variables, but it cannot model dynamics among states. On the other sides, Latent-Dynamic Conditional Random Field (LDCRF) can model the sub-structure of a state and learn dynamic among states [6]. The LDCRF model combines the strengths of CRF and HCRF. Furthermore, it can detect and recognize states from un-segment data (Figure1).

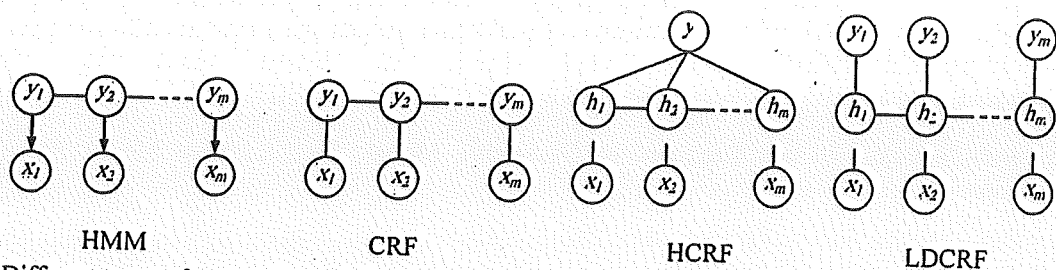


Figure 1. Different types of generative and discriminative models: HMM, CRF, HCRF and LDCRF. In these models, x_j refers to the j^{th} corresponding observation value, h_j is a hidden states that assigned to x_j . y_j is the label of x_j where the gray circles represent the observed variables.

The main contribution of this paper is to investigate the generative and discriminative models for recognizing the patterns of alphabets characters (A-Z) and numbers (0-9). The experimental results discover that LDCRF is the best in terms of results than CRF, HCRF and HMM. Additionally, LDCRF can automatically recognize hand patterns (i.e. gestures) with 98.05%. The rest of this paper is organized as follow; Section 2 reviews the basic classification generative and discriminative techniques. The main difference between generative and discriminative models is described in Section 3. Experimental results are described in Section 4. Section 5 concludes this paper.

2. Classification

In computer vision, a good choice for classification approaches helps the success of any system and makes it suitable for real-world applications. Classification of symbols in pattern recognition assigns them to respective classes. An application of hand gesture-based interaction with alphabet characters and numbers is implemented to demonstrate the coactions of suggested components and the effectiveness of gesture (i.e., hand motion) recognition. The isolated hand gesture is handled according to two different classification techniques: generative model such as HMM and discriminative model like CRF, to decide which one is the best in terms of performance. The following two subsections discuss how HMM and CRF are employed for the classification of alphabets character and numbers.

2.1 Generative Model: HMM

The most widely used recognition algorithm for pattern recognition is HMM [13, 14]. HMM is mathematical model of the stochastic process which generates a sequence of observations according to the previously stored information. In the Markov chain, every state of the model can only observe a single symbol. However, all states in Hidden Markov Model topology can observe one symbol out of a distinct gesture. The probability of observing a symbol for each state is stored in the observation probability distribution matrix. Furthermore, HMM states are called hidden for the following reasons. Firstly, the decision of observing a symbol represents the second process. Secondly, the emitter of an HMM only emits the observed symbol. Finally, the emitting states are unknown since the current states are based on the previous states. HMM have many advantages that are rich mathematical framework, powerful learning and decoding methods, good sequences handling capabilities, and flexible topology for the statistical phonology and the syntax. The disadvantages lie in the poor discrimination between the models and in unrealistic assumptions that must be made to construct the HMM theory, namely the independence of the successive feature frames (i.e. input vectors) and the first order Markov process [15].

2.1.1 Elements of HMM

A Hidden Markov Model can be symbolized with $\lambda = (A, B, \pi)$ and is characterized by the following elements;

- The set of states $S = (s_1, s_2, \dots, s_N)$, N represents the number of states in the model.
- An initial probability distribution for each state π such that:

$$\pi_i = p(s_i) \quad 1 \leq i \leq N \quad (1)$$

- An N -by- N transition matrix $A = \{a_{ij}\}$, which is given by:

$$a_{ij} = p(s_j | s_i), \quad 1 \leq i, j \leq N \quad (2)$$

where a_{ij} is the probability of the transition from state s_i at time t to s_j at time $t+1$. The sum of the entries in each row of matrix A must be 1 because it is the sum of the probabilities of making a transition from a given state to each other states.

$$\sum_j a_{ij} = 1 \quad (3)$$

- The set of possible emission (an observation) $O = \{o_1, o_2, \dots, o_T\}$ in which T is the length of gesture path.
- The set of discrete symbols $V = (v_1, v_2, \dots, v_M)$, where M represents the number of distinct observation symbols per state (i.e., the size of a codeword).
- An N -by- M observation matrix $B = \{b_{jm}\}$, where

$$b_{jm} = p(v_m | s_j) \quad 1 \leq j \leq N; \quad 1 \leq m \leq M \quad (4)$$

where b_{jm} gives the probability of emitting symbol v_m at state s_j . The sum of the entries in each row of matrix B must be 1 for the same previous reason. In short, a complete specification of the HMM contains two model parameters (N and M). Additionally, it also includes the observation symbols and the three probabilistic parameters A , B and π . Thus, a compact notation of HMM is as follows;

$$\lambda = p(\pi, A, B) \quad (5)$$

Here, λ refers to the parameters set of the model.

2.1.2 Three Basic Problems of HMM

Mathematically, three factors control the use of HMMs. These factors lie in their topologies, the selected features to be emitted and their observation probabilities. The feature selections are based on the observation task. Given the form of HMM, there are three basic problems that must be solved for the model to employ transitions and observation probabilities in a useful for real-world applications. The problems are;

Problem 1 "Evaluation Problem", Given the observation sequence $O = (o_1, o_2, o_3 \dots o_T)$ and a model parameter $\lambda = (A, B, \pi)$, how do we efficiently compute the probability of observation sequence given the model parameter $p(O|\lambda)$?

Problem 2 "Decoding Problem", Given the observation sequence $O = (o_1, o_2, o_3 \dots o_T)$, and the model parameter $\lambda = (A, B, \pi)$, how we choose a corresponding state sequence $Q = q_1 q_2 q_3 \dots q_T$ which is optimal in some meaningful sense.

Problem 3 "Estimation Problem", How do we adjust the model parameter $\lambda = (A, B, \pi)$ to maximize $p(O|\lambda)$?

2.1.3 Model Size

Before the HMM training starts, the size of HMM must be decided. How many states do we need? The number of states must be estimated by considering the complexity of the various patterns that HMM will be used to distinguish. In other words, the number of segmented parts in the graphical pattern is taken into consideration when we represented it. When the number of training data samples is insufficient, the use of excessive state numbers causes the over-fitting problem. The over-fitting occurs when HMM describes random error instead of the underlying relationship. Potential over-fitting problem does not only depend on the number of parameters and data, but also on the compatibility of model structure with the amount of model error and data shape. To avoid the problem of over-fitting, additional techniques (e.g. regularization, early stopping, cross-validation and etc.) are used when further training is not resulting in better generalization. For more details, the reader can refer to [1]. In addition, the discrimination power of the HMM is decreased when insufficient number of states is used because more than one segmented part of graphical pattern is modeled on one state. The number of states in our gesture recognition system is determined by mapping each straight-line segment into a single HMM state (Figure 2). To represent various graphical patterns, we must look at the possible patterns and estimate how many distinguishable segments are contained in a pattern. It may be a good idea to use different numbers of states in the different HMM, which used to represent separate classes of patterns. For example, to represent a graphical pattern '4', only five states are needed, whereas four states are required for a graphical pattern '3'.

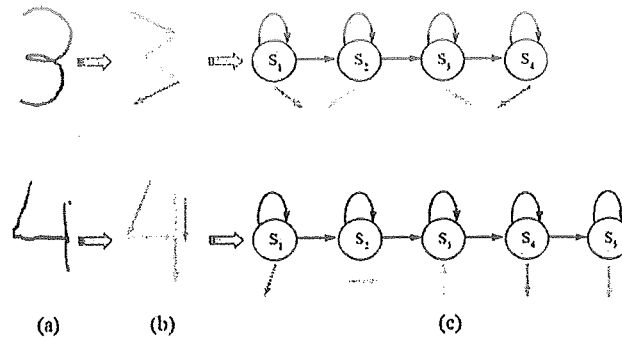


Figure 2. Straight-line segment for HMM topologies (a) the Gesture numbers from hand motion trajectory (b) Line segment of gesture number (c) Left-Right Banded (LRB) model with line segmented codeword.

2.1.4 Initializing Left-Right Banded Topology

Before starting the iterative Baum-Welch algorithm [13], the initial values of all parameters in the HMM must be assigned. There is only one general requirement; the initial model must indicate, somehow, what we want to represent different model states. However, this requirement has different consequences, depending on the type of HMM. In practice, the Left-Right Banded 1 (LRB) model is considered because each state in Ergodic topology has many transitions than Left-right (LR) and LRB topologies, so, the structure data can be easily lost. On the other hand, LRB topology has no backward transition, so the state index either increases or remains the same as time increases. In addition, LRB topology is more restricted than LR topology and simple for training data, which can match the data to the model [2]. An intuitively observation is that, a good initialization for HMMs parameters (A , B , π) achieves better results. Matrix A is the first parameter, where it is determined using the following Equation;

$$A = \begin{pmatrix} a_{11} & 1-a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & 1-a_{22} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \quad (6)$$

The diagonal elements a_{ii} of the transition matrix can be chosen to indicate approximately the average state durations d such that:

$$a_{ii} = 1 - \frac{1}{d} \quad (7)$$

And,

$$d = \frac{T}{N} \quad (8)$$

where T is the length of gesture path and N represents the number of states.

This is sufficient for an automatic training procedure in which State 1 is intended to represent the first part of the training data, State 2 the next part, etc. Therefore, all output probability distributions for different states can be initialized with the same parameters for all states. Consequently, the first step in Baum-welch iteration uses the training data to calculate more correct output probability parameters for each state. Since HMMs states are discrete, all elements of matrix B are initialized with the same value for all different states (Eq.9). Matrix B is an N -by- M observed symbols where b_{im} gives the probability of emitting symbol v_m in state i .

$$b_{im} = \frac{1}{M} \quad (9)$$

$$B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1M} \\ b_{21} & b_{22} & \dots & b_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ b_{N1} & b_{N2} & \dots & b_{NM} \end{pmatrix} = \begin{pmatrix} \frac{1}{M} & \frac{1}{M} & \dots & \frac{1}{M} \\ \frac{1}{M} & \frac{1}{M} & \dots & \frac{1}{M} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{M} & \frac{1}{M} & \dots & \frac{1}{M} \end{pmatrix} \quad (10)$$

where i, m run over the number of states and the number of discrete symbols respectively.

For each new time sample, the state can jump back by itself, or only to the nearest following state. Therefore, the initial probability vector π should be initialized as:

$$\pi = (1 \quad 0 \quad \dots \quad 0)^T \quad (11)$$

It is to ensure that it begins from the first state.

2.1.5 Termination of HMMs Training

The Baum-Welch training algorithm is very efficient. Often a good model is reached already after 5-10 iterations. The trained model must be flexible enough to correctly represent a new test sequence that never occurred during training. The training step is repeated until the change of transition and emission matrix converges. The convergence is satisfied if the change is less than 0.001 (i.e. tolerance $\varepsilon = 0.001$) as described in Eq. 12, or reaches to the maximum number of iterations (i.e. 500).

$$\sum_{i=1}^N \sum_{j=1}^N |\hat{a}_{ij} - a_{ij}| + \sum_{j=1}^N \sum_{m=1}^M |\hat{b}_{jm} - b_{jm}| < \varepsilon \quad (12)$$

The main motivation behind using tolerance is to control the number steps required by the Baum-welch algorithm in order to successfully execute its purpose. This algorithm is terminated if all of the following three quantities are less than the tolerance value. First, log-likelihood for a given observation sequence O is generated using the current estimated values of transition matrix A and observation matrix B . Second, change in the normalization of the transition matrix A .

At the end, change in the normalization of the observation matrix B . Note that, increasing tolerance reduces the number of steps to execute the Baum-welch algorithm before it was terminated. In fact, the maximum number of iterations controls the maximum number of steps to execute the algorithm. If the Baum-welch algorithm executes 500 iterations before reaches to the specified tolerance value, the termination is occurred with a warning. When this occurs, the value of maximum number of iterations should be increased so that the algorithm reaches to the desired tolerance before termination.

It is usually very difficult to provide sufficient amounts of training data. Therefore, some observation may never occur in the limited set of training data, although we may know that they might have occurred with some small probability. If a discrete HMM is trained on such data, the Baum-welch will assign zero observation probability to some elements of the observation probability matrix. In such case, a very small non-zero value may be assigned and re-normalization of the row matrix is required. A similar problem can occur with the transition probability matrix. For a left-right banded HMM we have intentionally defined many elements of the transition probability matrix exactly zero values. These elements still have zero values after the Baum-welch training, and should remain zero. Furthermore, the adjustment of HMM parameters is important after performing the training operation.

2.2 Discriminative models: CRF

CRF is undirected graphical models that were developed for labeling sequential data. Therefore, there is trade-off in the weights of occurrences of a feature value for each state [16]. The gesture recognition module matches the tested gesture against database of reference gestures, to classify which class it belongs to. Thereby, the hand gesture path is recognized corresponding to the maximal likelihood of all gestures (i.e. labels) accumulatively until it receives the gesture end signal. The maximal label of CRF model is the gesture whose observation probability is the largest among all the gestures labels (Figure. 1). CRF is differing than HMM in their conditional nature and the dependencies assumptions in their computations to ensure tractable inference. In addition, CRF overcomes the weakness of directed graphical models, which suffer from the bias problem as in Maximum Entropy Markov model (MEMM) [17, 18]. In our work, each label (state) corresponds to a hand gesture (e.g. characters from A to Z or numbers from 0 to 9). In addition, there is a trade-off for each label according to the weights of each feature function because CRF uses a single exponential distribution to model all reference labels of given observation [18]. The probability of label sequence y for a given observation sequence x is calculated by;

$$p(y|x, \theta) = \frac{1}{z(x, \theta)} \cdot \exp(\sum_{i=1}^n F_{\theta}(y_{i-1}, y_i, x, i)) \quad (13)$$

where $z(x; \theta)$ is the normalized factor given by;

$$Z(x, \theta) = \sum_y \exp(\sum_{i=1}^n F_{\theta}(y_{i-1}, y_i, x, i)) \quad (14)$$

where parameter $\theta = (\lambda_1, \lambda_2, \dots, \lambda_{N_f}; \mu_1, \mu_2, \dots, \mu_{N_g})$, N_f represents the number of transition feature function, N_g refers to the number of state feature function and n is the length of observation sequence x . F_{θ} is defined as follows;

$$F_{\theta}(y_{i-1}, y_i, x, i) = \sum_f \lambda_f t_f(y_{i-1}, y_i, x, i) + (\sum_g \mu_g s_g(y_i, x, i)) \quad (15)$$

where $t_f(y_{i-1}, y_i, x, i) \approx t_f(y_{i-1}, y_i, x)$ is a transition feature function of the entire observation sequence and labels at positions i and $i-1$ in the label sequence. $s_g(y_i, x, i) \approx s_g(y_i, x)$ refers to a state feature function of the label at position i and the observation sequence. λ_f and μ_g represent the weights of the transition and state feature functions respectively, which can be estimated from training data.

From the Eq. 13 and Eq. 15, the joint probability of a label sequence y given an observation sequence x can be written as follows:

$$P(y|x, \theta) = \frac{1}{z(x, \theta)} \cdot \exp(\sum_{i=1}^n \sum_f \lambda_f t_f(y_{i-1}, y_i, x, i) + \sum_{i=1}^n \sum_g \mu_g s_g(y_i, x, i)) \quad (16)$$

As CRF model is similar to HMM model in its characteristics, it is easy to build a CRF model by defining a single feature for each label observation pair (y_b, x) and label-label pair (y_a, y_b) according to the training data set as follow;

$$t_{y_a, y_b}(y_u, y_v, x) = \begin{cases} 1 & \text{if } y_u = y_a \text{ and } y_v = y_b \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

$$s_{y_b, x}(y_v, x_v) = f(x) = \begin{cases} 1 & \text{if } y_v = y_b \text{ and } x_v = x \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

Based on the foregoing mentioned, the parameters $\mu_{y_{b,x}}$ and $\lambda_{y_{a,y_b}}$ which corresponds to $S_{y_{b,x}}(y_v, x_v)$ and $t_{y_a, y_b}(y_u, y_v, x)$ features respectively are equivalent to the logarithms of the HMM observation and transition probabilities.

2.2.1 Learning CRF Model Parameter

The parameter estimation problem is to determine the parameter $\theta = (\lambda_1, \lambda_2, \dots, \lambda_{N_f}; \mu_1, \mu_2, \dots, \mu_{N_g})$ from the training data set $D = \{(x^{(j)}, y^{(j)})\}_{j=1}^{T_d}$, where $x^{(j)}$ is an observation sequence of training set, $y^{(j)}$ is the corresponding label sequence for observation sequence $x^{(j)}$ and T_d is the number of training sequences. The objective function to learn the parameter θ , which maximize the log-likelihood of the training data, is calculated by Eq. 19.

$$L(\theta) = \sum_{j=1}^{T_d} \log(p(y^{(j)}|x^{(j)}, \theta)) = \sum_{j=1}^{T_d} (\sum_{l=1}^n F_{\theta}(y_{l-1}^{(j)}, y_l^{(j)}, x_l^{(j)}, l) - \log z(x_l^{(j)}, \theta)) \quad (19)$$

Likelihood maximization can be performed using a gradient ascent method [19];

$$\frac{\partial L(\theta)}{\partial \theta} = \sum_{j=1}^{T_d} \left(\sum_{l=1}^n \frac{\partial F_{\theta}(y_{l-1}^{(j)}, y_l^{(j)}, x_l^{(j)}, l)}{\partial \theta} - \sum_x p(y|x^{(j)}) \sum_{l=1}^n \frac{\partial F_{\theta}(y_{l-1}, y_l, x_l, l)}{\partial \theta} \right) \quad (20)$$

In our work, a gradient ascent with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization technique is considered with 300 iteration to converge [19]. Also, HCRF and LDCRF models have a similar computational complexity to fully observable CRF.

2.2.2 Inference CRF

To compute the probability $p(y|x, \theta)$ of label sequence y for the given new observation sequence x , a set of matrices is computed [18]. To simplify some expressions, special starting y_0 and stopping y_{n+1} states are added. These states are dummy (i.e. observe no symbol and are passed without time delay). Suppose that $p(y|x, \theta)$ is given by Eq. 15. For each position i in the observation sequence, $M_i(x)$ is $|Y \times Y|$ matrix, which defined as follows;

$$M_i(y', y|x) = \exp(F_{\theta}(y', y, x, i)) \quad (21)$$

where $y = y_1, y_2, \dots, y_l$ represents a set of labels of the training data set. l refers to the number of the labels, and y', y are the labels at time i . Using this notation, the conditional probability of a label sequence y given the observation sequence x can be written as the product of the appropriate elements of the $n+1$ matrices for that pair of sequences (Eq. 22);

$$p(y|x, \theta) = \frac{1}{Z(x, \theta)} \cdot \prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|x) \quad (22)$$

Similarly, the normalization factor $Z(x, \theta)$ for observation sequence x is given by the (starting, stopping) entry of the product of all $M_i(x)$ matrices;

$$Z(x, \theta) = (\prod_{i=1}^{n+1} M_i(x))_{starting, stopping} \quad (23)$$

2.2.3 CRF with Hidden Variables

Other approaches including the hidden variables offer several advantages over previous CRF model. Although the CRF model the transition among gestures and overcome the weakness of directed graphical models which suffer from bias problem, it does not have the ability to learn the internal sub-structure of gesture sequences. Hidden Conditional Random Field (HCRF) is the extension of CRF, which incorporate hidden state variables to deal well with gesture sub-structure [18]. The main advantage of HCRF is to automatically model the local interconnection between labels (i.e. states) with hidden variables. However, it cannot model the dynamics among the states (Figure 1).

Latent-Dynamic Conditional Random Field (LDCRF) is considered as one of the approaches, which combine the advantages of CRF and HCRF by using both extrinsic dynamics and intrinsic sub-structure [18, 19]. The strategy of LDCRF is based on two main points. Firstly, they learn extrinsic dynamics by modeling the class labels. Secondly, they learn the intrinsic sub-structure of gesture sequence using intermediate hidden states. Thus, LDCRF model has the ability to overcome the main weaknesses of HCRF models (Figure 1). LDCRF model can be used to recognize the un-segmented sequences because they contain a class label per observation. Furthermore, LDCRF model can efficiently infer the gesture sequences during training and testing processes. HCRF models have only one label associated to each sequence while CRF and LDCRF have one label associated to each time sample in the sequence. As shown in Figure1, x_j refers to the j^{th} corresponding observation value, h_j is a hidden state that assigned to x_j . y_j represents the label of x_j where the gray circles refer to the observed variables.

2.2.4 Data Format of CRF

CRF and LDCRF models are applied to un-segmented sequences while HCRF should be applied to pre-segmented sequences (only one label per sequence). The data and the label files are encoded using Comma Separated Values (CSV) format according to HCRF library. The CRF formulation is implemented by extending the software of the library of Hidden-state Conditional Random Field [20]. This library implements three models: CRF, HCRF and LDCRF with C++ and Matlab languages.

Each file contains multiple matrices or vectors encoding the feature values (data files) or label values (label files). A data file contains multiple matrices, one for each sequence. For each matrix, the first line always contains two numbers: the number of rows and the number of columns. The number of rows for each matrix represents the number of features. All the matrices should have the same number of features. The number of columns for a specific matrix represents the number of time samples in the sequence. Since HCRF model has only one label associated to each sequence while CRF and LDCRF have one label associated to each time sample in the sequence, the HCRF library supports two file format for labels. For HCRF model, the label file contains one integer per line, representing the label for the specific sequence. For CRF and LDCRF models, the label file is encoded as a data file with matrix headers specifying the number of rows and columns but in this case the matrices always have one row. This row should have the same length as the corresponding sequence in the data file, with one label for each time sample.

3. Generative versus Discriminative models

The difference between HMM and CRF is that HMM is the generative models and define a joint probability distribution to solve a conditional problem, thus focusing on modeling the observation to compute the conditional probability. Moreover, one HMM is constructed per label (i.e. each alphabet character or number) where HMM assumes that all the observations are independent. CRF is undirected graphical model

and is developed for labeling sequential data. The key features of CRF than HMM are represented in their conditional nature and the dependencies assumptions of their computations to ensure tractable inference. In addition, CRF overcomes the weakness of directed graphical models which suffer from the bias problem as in MEMM [18]. Furthermore, CRF combines the strength of MEMM and HMM where they have all characteristics of the directed graphical models as in HMM. In addition, each label in CRF is employed as exponential model as in MEMM to conditional probabilities of the next label for a given current label. Additionally, CRF uses a single model for all alphabets and numbers.

4. Experimental Results

In our experimental results, the segmentation of the hand with complex background takes place using 3D depth map and color information over YC_bC_r color space, which is more robust to the illumination variation and partial occlusion. Gaussian Mixture Model (GMM) was considered where a large database of skin and non-skin pixel is used to train it. Moreover, morphological operations were used as a preprocessing, and Mean-shift algorithm in conjunction with Kalman filter [1] is to track the hand to generate the hand motion trajectory. Combined features of location, orientation and velocity for hand gestures are extracted and then, k -means clustering is employed for HMM, CRF, HCRF and LDCRF codeword. Our experiments are carried out on isolated gestures according to two different classification techniques: generative model and discriminative models. The following sections discuss the analysis of HMM and CRF results in details.

4.1 Data Set

The alphabet characters and numbers are classified using HMM, CRF, HCRF and LDCRF by the motion trajectory of single hand. A database is developed containing 2160 video samples for gesture symbols taken from three subjects on a set of 26 characters and 10 numbers. In other words, each isolated gesture is based on 60 video sequences where 42 video samples for training and 18 video samples for testing (In total, our database contains 1512 video samples for training and 648 video samples for testing). The sample test data is entirely different from the training data and is tested on *Intel(R) Core(TM)2 Duo CPU 2.2GHz PC with 4 GB of RAM*. The input images are captured by Bumblebee stereo camera system which has 6 mm focal length at 15FPS with 240×320 pixels image resolution, and Matlab implementation. Bumblebee camera is used for acquisition of 2D images. Therefore the databases are captured in IESK lab, Otto-von-Guericke-University Magdeburg, Germany.

4.2 Experimental Discussion

To handle isolated gesture, CRF, HCRF and LDCRF with different number of window sizes (W) ranging from 0 to 7 are applied and tested to decide the best in term of recognition results. A window size of zero means that the feature vector at the current frame is only used to construct the input feature while the window size of three means that the input feature vector at each frame consists of seven feature vectors which contain the current frame, three preceding frames and three future frames. In our application, the size of window is based on the complexity of each gesture as described in previous section. So, multiple experiments have been conducted with a variety of window size to empirically conclude the optimal outcome of the recognition system. Figure 3 shows the recognition rate of CRF, HCRF and LDCRF according to different window sizes for training and testing data. The recognition of hand gesture path using LDCRF is higher than CRF and HCRF. In addition, the yield of training data is higher than testing data in the proposed method. Furthermore,

the gesture recognition rate is initially improved as the window size increases but degrades as the window size further increases.

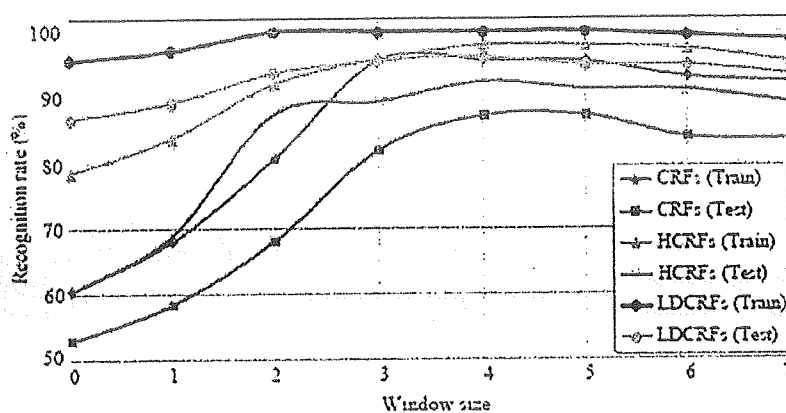


Figure 3. Recognition accuracy with different window sizes ranging from 0 to 7 for CRF, HCRF and LDCRF on training and testing data.

In HMM, we use a Left-Right Banded model based on Gaussian emission probabilities which have a full covariance matrix for each state. The HMM parameters (i.e. the emission probability and the state transition matrix) are learned from the same training data used by CRF. HMM is trained by Baum-welch algorithm while CRF is trained using Gradient ascent with the BFGS optimization technique. On a standard desktop PC, training process is more expensive for CRF, HCRF and LDCRF than HMM since the required time to model ranges from 20 Minutes to several hours and is based on observation window. On the contrary, the Inference (i.e. recognition) process is less costly and very fast for all models with Sequences of several frames (e.g. more than 80 frames in a sequence). The type of observed gesture is decided with HMM by Viterbi algorithm, frame by frame. As shown in Table 1, the overall recognition rates (the average of the training and the testing of recognition result) of HCRF at window size equal to 0 is higher than CRF. Also, in that case, the overall recognition rate achieved by HMM is 96.91%. Furthermore, HMM is the best in terms of results than CRF, HCRF and LDCRF at $W=0$.

Model type	Data set		Recognition result (%)		
	Training	Testing	Training	Testing	Overall
CRF	1512	648	60.34	52.78	56.56
HCRF	1512	648	78.55	60.34	69.45
LDCRF	1512	648	95.68	86.73	91.21
HMM	1512	648	99.07	94.75	96.91

Table 1. Results of gestures recognition at window size $W=0$

Whereas at window size equal to 4, LDCRF recognition rate is higher than HMM according to the training and the testing data (Figure 4). Our results show that the overall recognition rates are 91.52%, 95.28%, 96.94% and 98.05% for CRF, HCRF, HMM and LDCRF, respectively. The recognition ratio is the number of correctly recognized gestures over the number of input gestures (Eq. 24).

$$\text{Recognition ratio} = \frac{\text{number of recognised gestures}}{\text{number of tested gestures}} \times 100 \quad (24)$$

The high recognition rate achieved by the proposed system is due to the following reasons; 1) As a benefit of depth information, a high segmentation accuracy of the hand is achieved. 2) A set of feature candidates that optimally discriminate among the input patterns is elected. 3) A carefully experimental based selection of initialization parameters for training process. 4) HMM, CRF, HCRF and LDCRF classification techniques have the ability to efficiently alleviate spatio-temporal variabilities.

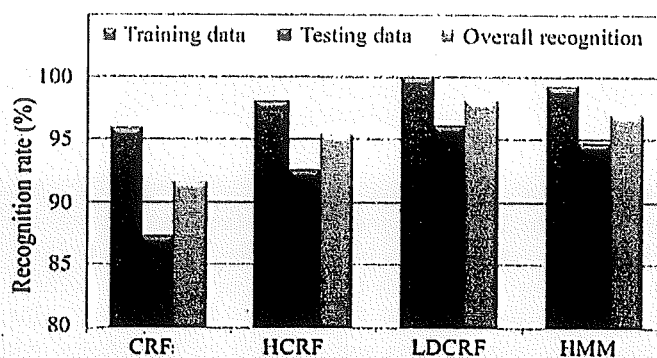


Figure 4. Results of gestures recognition using CRF, HCRF, LDCRF versus HMM at window size = 4

5 Conclusion

Experiments were carried out on isolated gestures according to two different classification techniques: a generative models such as HMM and discriminative models like CRF, HCRF and LDCRF. For discriminative models, CRF, HCRF and LDCRF with different number of window sizes ranging from 0 to 7 were applied and tested to decide the best among them. In contrast to generative and discriminative models, HMM was the best in terms of results than CRF, HCRF and LDCRF at window size = 0. Whereas at window size equal to 4, LDCRF recognition results were higher than HMM according to the training and the testing data. Our results showed that, the overall recognition rates were 91.52%, 95.28%, 96.94% and 98.05% for CRF, HCRF, HMM and LDCRF, respectively. It is noted that the proposed system achieves high recognition rate due to a high segmentation accuracy of hand. In addition, a good election for the set of feature candidates that optimally discriminate among input patterns. Also, a careful experimental based selection is required for initialization parameters of training process. Above all, HMM, CRF, HCRF and LDCRF classification techniques have the ability to efficiently alleviate spatio-temporal variabilities.

Acknowledgment

This work is supported by BMBF Bernstein-Group (FKZ: 01GQ0702), LSA grants (C4-NIMITEK 2, FKZ: UC4-3704M), Forschungspraemie (BMBF-Frderung, FKZ: 03FPB00213) and Transregional Collaborative

Research Centre SFB/TRR 62"Companion-Technology for Cognitive Technical Systems", funded by DFG. I am extremely grateful to Prof. Dr. Halle Ferjani, Faculty of Science, Tanta University, Egypt for guiding my work and supporting me in several directions.

Reference

- [1] M. Elmezain, A. Al-Hamadi and B. Michaelis, "Real-Time Capable System for Hand Gesture Recognition Using Hidden Markov Models in Stereo Color Image Sequences". Journal of WSCG, Vol. 16, No. 1, pp. 65-72, 2008.
- [2] M. Yang, N. Ahuja and M. Tabb, "Extraction of 2D Motion Trajectories and its Application to Hand Gesture Recognition". IEEE Trans. on PAMI, Vol. 24, No. 8, pp. 1061-1074, 2002.
- [3] H. Yang, S. Sclaroff and S. Lee, "Sign Language Spotting with a Threshold Model Based on Conditional Random Fields". IEEE Trans. on PAMI, Vol. 31(7), pp. 1264-1277, 2009.
- [4] C. Sminchisescu, A. Kananujia and D. Metaxas, "Conditional Models for Contextual Human Motion Recognition". Journal of CVIU, Vol.104, No. 2, pp. 210-220, 2006.
- [5] J. Lafferty, A. McCallum and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling sequence Data", Conf. on ICML, pp. 282-289, 2001.
- [6] L. P. Morency, A. Quattoni and T. Darrell, "Latent-Dynamic Discriminative Models for Continuous Gesture Recognition". IEEE Conf. on CVPR, pp. 1-8, 2007.
- [7] X. Deyou, "A Network Approach for Hand Gesture Recognition in Virtual Reality Driving Training System of SPG". Conf. on ICPR, pp.519-522, 2006.
- [8] M. Elmezain, A. Al-Hamadi and B. Michaelis, "Spatio- Temporal Feature Extraction-Based Hand Gesture Recognition for Isolated American Sign Language and Arabic Numbers". IEEE Symposium on ISPA, pp. 254-259, 2009.
- [9] R. R. Lawrence, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceeding of the IEEE, Vol.77(2), pp. 257-286, 1989.
- [10] K. Takahashi, S. Sexi and R. Oka, "Spotting Recognition of Human Gestures From Motion Images". In Technical Report IE92-134, pp.9-16; 1992.
- [11] A. Gunawardana, M. Mahajan, A. Acero and J. C .Platt, "Hidden Conditional Random Fields for Phone Classification". Proceeding of European Conf. on Speech Communication and technology, pp. 1117-1120, 2005.
- [12] A. Quattoni, S. Wang, L. P. Morency, M. Collins and T. Darrell, "Hidden Conditional Random Fields". IEEE Trans. on PAMI, Vol. 29, No.10, pp. 1848-1852, 2007.
- [13] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE, Vol. 77, No. 2, pp. 257-286, 1989.
- [14] X. D. Huang, Y. Ariki, and M. Jack, " Hidden Markov Models for Speech Recognition", Edinburgh University Press, 1990.
- [15] S. Goronzy, " Robust Adaptation to Non-Native Accents in Automatic Speech Recognition", Lecture Notes in Computer Sciences, Springer, ISBN-13: 978- 540003250, 2002.
- [16] M. Elmezain, A. Al-Hamadi, and B. Michaelis, "Discriminative Models-Based Hand Gesture Recognition", International Conference on Machine Vision, pp. 123-127, 2009.
- [17] A. McCallum, D. Freitag, and F. Pereira, "Maximum Entropy Markov Models for Information Extraction and Segmentation". International Conference on Machine Learning, pp. 591-598, 2000.
- [18] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling sequence Data". International Conference on ICML, pp. 282-289, 2001.
- [19] A. McCallum, "Efficiently Inducing Features of Conditional Random Fields". Conf. on Uncertainty in AI, 2003.
- [20] L.-P. Morency, A. Quattoni, C. M. Christoudias, and S. Wang, "Hidden-state Conditional Random Field Library". Version 1.3c, <http://pt.sourceforge.jp/projects/sfnet hcrf/>," 2008.

1. The first part of the document is a letter from the President of the United States to the Congress, dated January 1, 1861. It is a very important document, as it sets out the President's policy for the new year. The letter is written in a very formal and dignified style, and it is a very good example of the President's power and authority.

2. The second part of the document is a letter from the President to the Congress, dated January 1, 1861. It is a very important document, as it sets out the President's policy for the new year. The letter is written in a very formal and dignified style, and it is a very good example of the President's power and authority.

3. The third part of the document is a letter from the President to the Congress, dated January 1, 1861. It is a very important document, as it sets out the President's policy for the new year. The letter is written in a very formal and dignified style, and it is a very good example of the President's power and authority.

4. The fourth part of the document is a letter from the President to the Congress, dated January 1, 1861. It is a very important document, as it sets out the President's policy for the new year. The letter is written in a very formal and dignified style, and it is a very good example of the President's power and authority.

5. The fifth part of the document is a letter from the President to the Congress, dated January 1, 1861. It is a very important document, as it sets out the President's policy for the new year. The letter is written in a very formal and dignified style, and it is a very good example of the President's power and authority.

Arabic Named Entity Recognition for a New Corpus of Egyptian People's Assembly

Ahmed Montasser H. Farag

Abstract

Named Entity Recognition (NER) task has become essential to improve the performance of many NLP tasks. Its aim is to endeavor a solution to increase the accuracy of extracted named entities' identification. This paper presents the first step to extract useful information for a researcher who is interested in Egyptian People's Assembly by creating a new corpus of the Egyptian People's Assembly and presenting a novel solution for Arabic Named Entity Recognition (ANER). The solution uses a Conditional Random Field (CRF) sequence labeling model by training it on mixing feature, morphological, gazetteers, and using character n-gram of leading and trailing letters in words. The results in this paper appear that the F-measure of mixing features running on the datasets of the in Egyptian People's Assembly is the better F-measure.

Keywords

Arabic Named Entity Recognition, Named Entity Recognition, Conditional Random Field, Egyptian People's Assembly.

1. Introduction

Named Entity Recognition (NER) has earned an important place in Natural Language Processing (NLP) as an enabling process for other tasks. NER task has been used to evolve many Natural Language Processing (NLP) subtasks, such as Information Retrieval, Question Answering, and Information Extraction [1]. NER is concerned with identifying the named entities (NE's) from the documents such as persons, locations, and organizations. For example, in the next sentence "Dalia Ibraheem, CEO of Nahdet Misr Company in Egypt" Dalia Ibraheem, Nahdet Misr Company, and Egypt would be identified as person, organization, and location respectively. Arabic language is a high complex language, and has a lot of challenges to extract NER from it. Some of these challenges are:

- Arabic is not case sensitive language; lack capitalization of proper nouns.
- Arabic is a high inflectional language; conjunctions, prepositions, possessive pronouns, and determiners are attached to words as prefixes or suffixes. For example, the Arabic word "وبحسناتهم" "wbHsnAthm" ¹ which means in English "and by their virtues".

¹ Arabic letters are presented using the Buckwalter transliteration scheme

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

- Arabic words have different meaning. For example, “رمضان” “Ramadan” in Arabic may be used as person name, month [2].

The paper introduces a new corpus for Egyptian People’s Assembly data, it took great effort to search for and collect data from the Parliament sessions’ minutes of the year 2006. This paper depended on reliable data sources like: “Al Ahram and Al youm7 newspapers. Also, we collected data from Delta Misr website. We made a lot of preprocessors on these documents to make them suitable for our work. We used a Conditional Random Field (CRF) sequence labeling model by training it on mixing features as morphological, gazetteers, and using character n-gram of leading and trailing letters in words that help handle prefix and suffix attachment. The contributions of this paper are as follows:

Present a new corpus for Egyptian People’s Assembly, collected from the minutes of Egyptian People’s Assembly in 2006, AlAhram News, Alyoum7 news, and DeltaMisr news.

Using mixing features like morphological, gazetteers, and using character n-gram of leading and trailing letters in words, which captures valuable clues that would indicate or counter-indicate the presence of NE’s.

Conditional Random Fields (CRF) sequence labeling was used in identifying NE’s and the experiments were performed on the new corpus of Egyptian People’s Assembly.

The paper is organized as follows: Section 2 surveys work on Arabic NER done before; Section 3 introduces the proposed feature and explains their use; Section 4 describes experimental setup; Section 5 the experimental results; and Section 6 concludes of the paper.

2. Related Work

The most successful language independent approaches to English NER are systems that employ Maximum Entropy (ME) techniques in a supervised setting [3].

(Tran et al., 2007) show that using a Support Vector Machine (SVM) approach outperforms (F=87.75) using CRF (F=86.48) on the NER task in Vietnamese [4].

For Arabic NER, (Benajiba et al., 2007) shows that using a basic ME approach yields F=55.23 [5]. Then they are followed up with further work in (Benajiba and Rosso, 2007), they model the problem as a two-step classification approach applying ME, separating the NE boundary detection from the NE classification. This modification shows performance improvement in yielding an F=65.91. None of these studies included Arabic specific features; all the features used were language independent [6].

In (Benajiba and Rosso, 2008) report using lexical and morphological features in a single step model using CRF which resulted in significant improvement over state of the art for Arabic

NER, yielding $F=79.21$. However, the data that was used in these evaluation sets were not standard sets [7].

Most recently, (Farber et al., 2008) have explored using a structured perception based model that employs Arabic morphological features. Their system benefits from the basic POS tag (15 tags) information and the corresponding capitalization information on the gloss corresponding to the Arabic word. Exploiting this information yields a significant improvement in recall of 7% and an overall $F=69.6$ on the ACE2005 data set [8].

In Benajiba et al. (2008), they examined the same feature set on the Automatic Content Extraction (ACE) datasets using CRF sequence labeling and Support Vector Machine (SVM) classifier. They did not report per category F-measure, but they reported overall 81%, 75%, and 78% macro-average F-measure for broadcast news and newswire on the ACE 2003, 2004, and 2005 datasets respectively [1].

Mayfield et al. (2003) used thousands of language independent features such character n-grams, capitalization, word length, and position in a sentence, along with language dependent features such as POS tags and BP chunking. For English, they reported 89%, 79%, and 91% F-measure for location, organization, and persons respectively [9].

K. Darwish et al. (2010) used CRF sequence labeling model trained by features that primarily use character n-gram of leading and trailing letters in words and word n-grams. That features help overcome some of the morphological and orthographic complexities of Arabic. For Arabic, They reported their paper are lower by 2 F-measure points for locations, but are better by 8 points for organizations and 9 points for persons [2].

Aly Fahmy et al. (2010) used a novel solution for Arabic Named Entity Recognition (ANER) problem. The solution is an integration approach between two machine learning techniques, namely bootstrapping semi-supervised pattern recognition and Conditional Random Fields (CRF) classifier as a supervised technique. For Arabic, they reported overall 68 %, 88%, 70% F-measure for person, location, organization respectively [10].

Though, the use of CRF sequence labeling has been increasing over the past few years [11] [12] with good success [7].

3. ANER feature

In this paper, a Conditional Random Field (CRF) labeling was used. It is a discriminative probabilistic model which is used for segmenting and labeling the sequential data. It is a generalization of Hidden Markov Model in which its combine HMM-like generative power with classifier- likes discrimination [13] [14]. Its undirected graph consists of nodes to represent

the label sequence y corresponding to the sequence x . The aim of CRF model is to find y that maximizes $p(y | x)$ for that sequence.

$$p(y|x) = \frac{1}{z(x)} * e^{\sum_t \sum_k \lambda_k f_k(y_{t-1}, y_t, x)}$$

Where $z(x) = \sum_{y \in Y} e^{\sum_t \sum_k \lambda_k f_k(y_{t-1}, y_t, x)}$; λ_k is the weight of f_k .

As a CRF make a decision on the label assigned to a word, take in their account the previous and succeeding words. We also used in this paper An AMIRA toolkit was used to tag words with POS, and Base Phrase Chunk (BPC). The AMIRA toolkit includes a clitic tokenizer (TOK), part of speech tagger (POS) and base phrase chunker (BPC) - shallow syntactic parser. The technology of AMIRA is based on supervised learning with no explicit dependence on explicit modeling or knowledge of deep morphology. AMIRA is based on using a unified framework casting each of the component problems as a classification task. The underlying technology employs Support Vector Machines in a sequence modeling framework using the YAMCHA toolkit [15]. Also we used the ANERGazet² is a collection of 3 Gazetteers, (i) Locations: a Gazetteer containing names of continents, countries, cities, etc.; (ii) People: a Gazetteer containing names of people recollected manually from different Arabic websites; and finally (iii) Organizations: containing names of Organizations like companies, football teams, etc.

The features used were as follows:

- Word (**WF**) is the word itself.
- Part of Speech (**POS**) is tagging of the word. The POS tagging system in AMIRA toolkit optionally produces the PATB standard tag set of 25 tags (RTS) as well as an extended tag set (ERTS) that encodes some overt morphological markers including gender, number and definiteness [16]. ERTS has 72 tags.
- Base Phrase Chunks (**BPC**) was atomic part of the sentence.
- Gazetteers (**GAZ**) are a binary feature to present the existence of the word in the gazetteers.
- Semantic fields (**SF**) feature presents some word semantic fields identification. This feature has important effect in ANER. For example, "عبد" "Abd", "أبو" "Abo", are very repetitive prefix in Arabic person names.

- Leading and trailing character bigrams (6bi), for a given word composed of the letter sequence l_0^n , where l_0 and l_n are a start and end word markers respectively, the first three bigrams (l_0^1 , l_1^2 , and l_2^3) and last three bigrams (l_{n-3}^{n-2} , l_{n-2}^{n-1} , and l_{n-1}^n) were used as features.

Using leading and trailing character bigrams of a word was an attempt to account for capture surface clues that would indicate the presence of a NE or not. For example, plural forms of common words in Arabic are often obtained by attaching the suffixes wn (ون) or yn (ين) for masculine nouns and At (ات) for feminine nouns. Presence of such plural form markers would generally indicate a plural noun, but would counter-indicate a NE. Also, verbs in present tense start with the letters A (ا), t (ت), y (ي), and n (ن). These would contribute to concluding that a word may not be a NE. Further, coordinate conjunctions, such as f (ف) and w (و), and prepositions, such as b (ب), k (ك), and l (ل), composed of single letters are often attached as prefixes to words. Accounting for them may help overcome some of the problems associated with not stemming. Further, the determiner Al (ال) may be a good indicator for proper nouns particularly in the case of organizations.

- Leading and trailing character trigrams (6tri). For a given word composed of the letter sequence l_0^n , where l_0 and l_n are a start and end word markers respectively, the first three trigrams (l_0^2 , l_1^3 , and l_2^4) and last three trigrams (l_{n-4}^{n-2} , l_{n-3}^{n-1} , and l_{n-2}^n) were used as features. The rationale for using these features is very similar to that of using character bigrams. The added value of using character trigrams is that they would allow for the capture of combinations of prefixes and suffixes. For example, a word may begin with the prefixes $w+Al$ (وال), which are a coordinating conjunction and determiner respectively.
- Leading and trailing character 4-grams (6quad). For a given word composed of the letter sequence l_0^n , where l_0 and l_n are a start and end word markers respectively, the first three 4 grams (l_0^3 , l_1^4 , and l_2^5) and last three 4 grams (l_{n-5}^{n-2} , l_{n-4}^{n-1} , and l_{n-3}^n) were used as features. Similar to leading and trailing trigrams, these features can capture combinations of prefixes and suffixes [2].

4. Experimental setup

4.1. Dataset

In this paper, the NE's of interest were persons, locations, and organizations only. Three datasets were used for the work in this paper. The first dataset we present in this paper is composed of the Lecturer of Egyptian People's Assembly in 2006 totaling more than 200,000 words. The number of NE's is as follows:

Locations (LOC)	674
Organizations (ORG)	1019
Persons (PERS)	1038

The second is composed of AlAhram news of Egyptian People's Assembly from 2001 to 2004 totaling more than 200,000 words. The number of NE's is as follows:

Locations (LOC)	480
Organizations (ORG)	809
Persons (PERS)	1226

The third is composed of Alyoum7 news of Egyptian People's Assembly in 2010, DeltaMisr news of Egyptian People's Assembly in 2010, and AlAhram news of Egyptian People's Assembly from 2007 to 2010. Egyptian People's Assembly dataset tags contain only the base tags (LOC, ORG, and PERS). The total number of words in the collection is 120,000 words.

The number of NE's is as follows:

Locations (LOC)	325
Organizations (ORG)	714
Persons (PERS)	861

Each collection was 80/20 split for training and testing.

4.2. Data Processing and Sequence Labeling

We make in this paper a lot of data preprocessing to create the corpus. First, we used "Parser News" which is a program in house to convert different kinds of HTML's collected from different web sites news, to row text. Second, we used AMIRA toolkit to tag the text by Part of Speech (POS) tag and Base Phrase Chunk (BPC) tag. Also, The coordinating conjunctions w (و), l (ل), b (ب), and f (ف), which always appear as the first prefixes in a word, were stemmed. w , l , b and f were stemmed using AMIRA toolkit. Third, we reformat the output files from AMIRA toolkit to be columns separated by tab. Finally, we were tagging the corpus manually.

4.3. Evaluation

The figures of merit for evaluation were precision, recall, and F-measure ($\beta=1$), with evaluation being conducted at the phrase level. A set of 8 experiments are being reported, using the short names of features (bolded after feature names in section 3), and the experiments were as follows:

- WF
- WF_POS
- WF_POS_BPC
- WF_POS_BPC_GAZ
- WF_POS_BPC_GAZ_SF
- WF_POS_BPC_GAZ_SF_6bi
- WF_POS_BPC_GAZ_SF_6bi_6tri
- WF_POS_BPC_GAZ_SF_6bi_6tri_6quad

5. Experimental results

Table 1 lists the results. Tables 2, 3, and 4 report the best obtained results. The results include precision (P), recall (R), and F-measure (F) for NE's of types location (LOC), organization (ORG), and person (PER). The best results for P, R, and F are bolded in the tables.

Run Name	Type	Lecturer of EPA			Ahram			News		
		P	R	F	P	R	F	P	R	F
WF	Per	98	89	93	99	88	93	98	83	90
	Loc	97	53	69	91	61	73	91	59	72
	Org	90	43	58	91	67	77	93	72	81
WF_POS	Per	99	91	95	99	91	95	98	85	91
	Loc	97	55	70	92	63	75	91	61	73
	Org	90	48	63	90	68	77	93	72	81
WF_POS_BP C	Per	99	91	95	99	91	95	97	85	91
	Loc	97	54	69	92	63	75	91	62	74
	Org	89	48	62	89	67	76	92	74	82
WF_POS_BP C_GAZ	Per	98	93	95	98	92	95	97	88	92
	Loc	96	65	78	90	70	79	89	69	78
	Org	89	48	62	89	66	76	92	74	82
WF_POS_BP C_GAZ_SF	Per	98	95	96	99	93	96	97	88	92
	Loc	96	65	78	89	73	80	89	74	81
	Org	83	52	64	88	70	78	91	78	84

WF_POS_BP C_GAZ_SF_6 bi	Per	98	95	96	98	94	96	98	91	94
	Loc	97	62	76	91	75	82	87	75	81
	Org	83	57	68	88	70	78	90	78	84
WF_POS_BP C_GAZ_SF_6 bi_6ti	Per	98	95	96	98	94	96	98	91	94
	Loc	97	61	75	92	75	83	88	76	82
	Org	85	59	70	88	72	79	91	78	84
WF_POS_BP C_GAZ_SF_6 bi_6ti_6quad	Per	98	95	96	98	94	96	98	91	94
	Loc	98	60	74	92	76	83	89	77	83
	Org	86	61	71	89	72	80	92	78	84

Table 1: NER results for Lecturer of EPA, Ahram News, and News.

	P	R	F
Per	98	95	96
Loc	98	60	74
Org	86	61	71
Avg.	94	72	80

Table 2: Best results on Lecturer of EPA dataset

Run Name: (WF_POS_BPC_GAZ_SF_6bi_6ti_6quad)

	P	R	F
Per	98	94	96
Loc	92	76	83
Org	89	72	80
Avg.	93	81	86

Table 3: Best results on Ahram News dataset

Run Name: (WF_POS_BPC_GAZ_SF_6bi_6ti_6quad)

	P	R	F
Per	98	91	94
Loc	89	77	83
Org	92	78	84
Avg.	93	82	87

Table 4: Best results on News dataset

Run Name: (WF_POS_BPC_GAZ_SF_6bi_6ti_6quad)

The **WF_POS_BPC_GAZ_SF_6bi_6ti_6quad** run produced nearly the best F-measure for both datasets, with extra features improving overall F-measure by at Least 1 point.

The notable part of this work is that using of mixing feature, morphological, gazetteers, and using character n-gram of leading and trailing letters in words.

6. Conclusion and future work

This paper presents the first step to extract useful information for the researcher who has interest in the Egyptian People's Assembly by creating a new corpus of Egyptian People's Assembly and presenting a novel solution for ANER. The solution uses a CRF sequence labeling model by training it on mixing feature, morphological, semantic fields, gazetteers, and using character n-gram of leading and trailing letters in words.

For future work, the author would like to use this work to extract more useful information from the news of Egyptian People's Assembly, and to examine the effectiveness of the proposed feature set on other datasets.

References

- [1] Y. Benajiba, M. Diab, and P. Rosso. 2008. *Arabic Named Entity Recognition using Optimized Feature Sets*. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 284–293, Honolulu, October 2008.
- [2] K. Darwish, and A. Abdul-Hamid. 2010. *Simplified Feature Set for Arabic Named Entity Recognition*. Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 110–115, Uppsala, Sweden, 16 July 2010.
- [3] O. Bender, F. Och, and H. Ney. 2003. *Maximum Entropy Models For Named Entity Recognition*. In Proceedings of CoNLL-2003. Edmonton, Canada.
- [4] Q. Tran T., T. Pham X. T., Q. Ngo H., D. Dinh, and N. Collier. 2007. *Named Entity Recognition in Vietnamese documents*. Progress in Informatics Journal. 2007.
- [5] Y. Benajiba, P. Rosso and J. M. Benedí. 2007. *ANERsys: An Arabic Named Entity Recognition system based on Maximum Entropy*. In Proc. of CICLing-2007, Springer-Verlag, LNCS(4394), pp.143-153.
- [6] Y. Benajiba and P. Rosso. 2007. *ANERsys 2.0: Conquering the NER task for the Arabic language by combining the Maximum Entropy with POS-tag information*. In Proc. of Workshop on Natural Language- Independent Engineering, IICAI-2007.
- [7] Y. Benajiba and P. Rosso. 2008. *Arabic Named Entity Recognition using Conditional Random Fields*. In Proc. of Workshop on HLT & NLP within the Arabic World, LREC'08.
- [8] B. Farber, D. Freitag, N. Habash, and O. Rambow. 2008. *Improving NER in Arabic Using a Morphological Tagger*. In Proc. of LREC'08.
- [9] J. Mayfield, P. McNamee, and C. Piatko. 2003. *Named Entity Recognition using Hundreds of Thousands of Features*. HLT-NAACL 2003- Volume 4, 2003.
- [10] A. Fahmy, S. AbdelRahman, M. Elarnaoty, and M. Magdy. 2010. *Integrated Machine Learning Techniques for Arabic Named Entity Recognition*. IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 3, July 2010.
- [11] A. McCallum and W. Li. 2003. *Early Results for Named Entity Recognition with Conditional Random Fields, Features Induction and Web- Enhanced Lexicons*. In Proc. Conference on Computational Natural Language Learning.

[12] D. Nadeau and S. Sekine. 2009. *A survey of named entity recognition and classification*. Named entities: recognition, classification and use, ed. S. Sekine and E. Ranchhod, John Benjamins Publishing Company.

[13] J. Lafferty, A. McCallum, and F. Pereira. 2001. *Conditional random fields: Probabilistic models for segmenting and labeling sequence data*, In Proc. Of ICML, pp.282-289, 2001.

[14] F. Sha and F. Pereira. 2003. *Shallow parsing with conditional random fields*, In Proc. Of HLT/NAACL-2003.

[15] M. Diab. 2009. *Second Generation Tools (AMIRA 2.0): Fast and Robust Tokenization, POS tagging, and Base Phrase Chunking*. Proceedings of the Second International Conference on Arabic Language Resources and Tools, 2009.

[16] Mona Diab. 2007b. Towards an optimal pos tag set for modern standard arabic processing. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria.

Applied Data Mining Classification Techniques For Diagnosing Breast Cancer Grades in Egypt

Rabab Emad Mohammed

Abstract

This paper focus on mining data to extract a classification criteria for diagnosing breast cancer grades [benign grade (G0), Malignant grades (G1, G2, G3)] that applied in National Cancer Institute dataset (Egypt) using a pathological attributes such as [Clump Thickness, Uniformity of Cell Size, Bare Nuclei, Mitoses, age, location], searching for good patterns in data using data cleaning, feature extraction, statistical and mining techniques, and data mining tool [DTREG]. An Applied four models are [Decision Tree, Multilayer Perceptron, Support Vector Machine, Gene Expression Programming] to determine which of them has perfect results in diagnosing breast cancer grade. In modeling phase, Variance Metrics such as [CV, NMSE, RMSE, MSE and MAE] and analysis time are used to evaluate the performance of the applied techniques. Results showed that Decision Tree is a best model with lowest variance error metrics and lowest analysis time in validation phase. Measuring the importance of parameters determine the most important parameters in diagnosing tumor grade. Analysis of Results showed that the highly importance parameters after applying four models for classifying tumor grade are [ucellsize, ucellshape, bnuclei, mgadhesion, bchromatin, normnuclei].

Keywords:

Breast cancer; Diagnosis, Data Mining, Classification, Decision Tree, Multilayer Perceptron, Support Vector Machine, Gene Expression Programming, DTREG, benign, malignant, CRISP-DM.

I. INTRODUCTION

Breast cancer is the most common cancer among women worldwide [1]. It refers to a malignancy in women that arises from the terminal ductal-lobular units of epithelial tissue, which in the mature breast represent 10% of the total volume [2]. It was estimated that during 2002, more than 1,050,000 cases developed worldwide annually with nearly 580,000 incident cases occurred in developed countries and the remainder in developing countries [3]. It is also the most important cause of neoplastic deaths among women. The estimated number of deaths in 2002 was 410,000 worldwide. In Egypt, contributing for 30% of all female cancers. It affects one in 14 women during their life time, almost one every three affected women will die of the disease. It constitutes 33% of all female cancers at the National Cancer Institute, Cairo University [4].

Sadat Academy For Management Sciences

Studies in (Damietta 2009 and Aswan 2008) mentioned that breast cancer was the most frequent cancer among females .The [229 (Damietta, 2009) and 241 (Aswan, 2008)] registered cases represented [32.9% (Damietta,2009) and 38.6% (Aswan, 2008)] of all newly diagnosed female cancers. The number of cases with available microscopic diagnosis was [209 (91.3%) (Damietta, 2009) and 173 (71.8%) (Aswan, 2008)] . Tumor grade was mentioned in [(69.9% = 160/229) (Damietta, 2009) and (66% = 159/241) (Aswan, 2008)] of registry records. The majority of cases were grade II and III [(76.3% and 21.9%) (Damietta, 2009) and (81.1% and 15.1%) (Aswan, 2008)]. Grade I was infrequent; [(1.9%) (Damietta, 2009) and (3.7%) (Aswan, 2008)]. [5] [6]

The problem associated with classifying the breast tumors whether malignant or benign is that each type have share structural similarities and it becomes an extremely uninteresting and time consuming task to manually differentiate them. So, a need for creating an automated system to provide a faster and more reliable method of diagnosis will appear.

The aim of this paper is creating a Classification breast cancer grading model by identifying characteristics or patterns that are associated with a high or low importance in applied two cases, and then identifying the highly shared importance patterns in each model.

II. BACKGROUND

1. Breast Cancer:

Breast cancer is the form of cancer that either originates in the breast or is primarily present in the breast cells [7]. The disease occurs mostly in women but a small population of men is also affected by it. Breast cancer is the most common form of cancer amongst the female population as well as the most common cause of cancer deaths after lung cancer.

Grading compares the appearance of the breast cancer cells to the appearance of normal breast tissue. Normal cells in an organ like the breast become differentiated, meaning that they take on specific shapes and forms that reflect their function as part of that organ. Cancerous cells lose that differentiation. In cancer, the cells that would normally line up in an orderly way to make up the milk ducts become disorganized. Cell division becomes uncontrolled. Cell nuclei become less uniform. Pathologists describe cells as well differentiated (low grade), moderately differentiated (intermediate grade), and poorly differentiated (high grade) as the cells progressively lose the features seen in normal breast cells [8].

2. Knowledge Discovery and Data Mining

The terms Knowledge Discovery in Databases (KDD) and Data Mining are often used interchangeably. KDD is the process of turning the low-level data into high-level knowledge. Hence, KDD refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and KDD are often treated as

equivalent words but in real data mining is an important step in the KDD process. The following fig. 1 shows data mining as a step in an iterative knowledge discovery process.

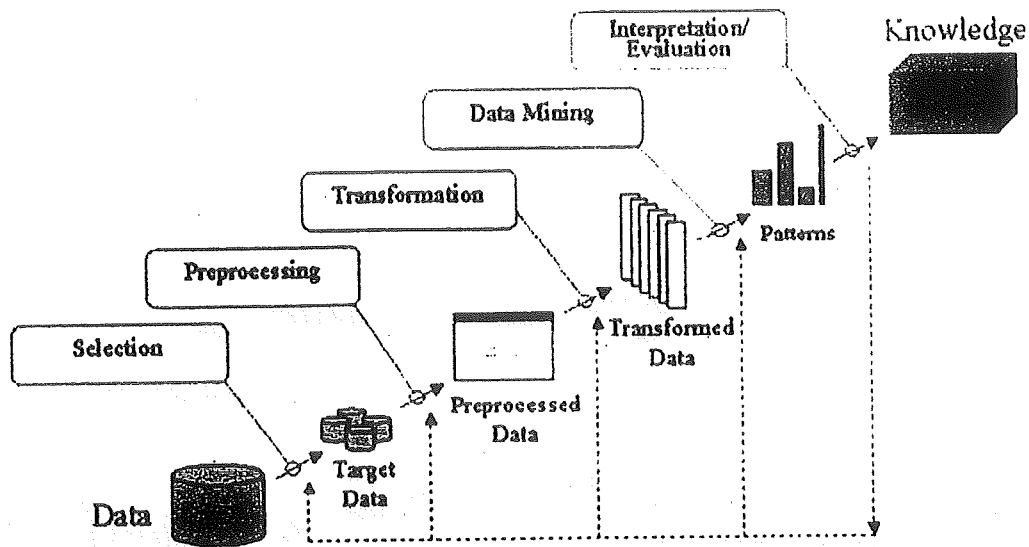


Figure 1: KDD Process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge [9]. The iterative process consists of the following steps:

- (1) **Data cleaning:** also known as data cleansing it is a phase in which noise data and irrelevant data are removed from the collection.
- (2) **Data integration:** at this stage, multiple data sources, often heterogeneous, may be combined in a common source.
- (3) **Data selection:** at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- (4) **Data transformation:** also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- (5) **Data mining:** it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- (6) **Pattern evaluation:** this step, strictly interesting patterns representing knowledge are identified based on given measures.
- (7) **Knowledge representation:** is the final phase in which the discovered knowledge is visually represented to the user. In this step visualization techniques are used to help users understand and interpret the data mining results.

3. Data Mining Concept and Process:

"Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques." [10] There are other definitions:

"Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" [11].

"Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" [12].

CRISP-DM is a widely accepted methodology for representing a data mining process [13].

Figure 2: shows the CRISP-DM steps.

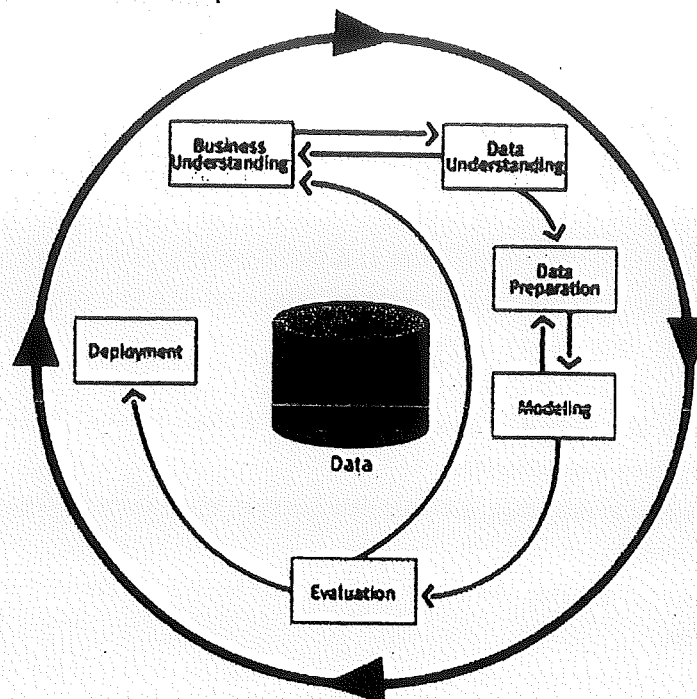


Figure 2: CRISP-DM steps

1. **Business Understanding:** Understand the project objectives and requirements from a business perspective, and then convert this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.
2. **Data Understanding:** Start by collecting data, then get familiar with the data, to identify data quality problems, to discover first insights into the data, or to detect interesting subsets to form hypotheses about hidden information.

3. **Data Preparation:** Includes all activities required to construct the final data set (data that will be fed into the modeling tool) from the initial raw data. Tasks include table, case, and attribute selection as well as transformation and cleaning of data for modeling tools.
4. **Modeling:** Select and apply a variety of modeling techniques, and calibrate tool parameters to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.
5. **Evaluation:** Thoroughly evaluate the model, and review the steps executed to construct the model, to be certain it properly achieves the business objectives. Determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results is reached.
6. **Deployment:** Organize and present the results of data mining. Deployment can be as simple as generating a report or as complex as implementing a repeatable data mining process.

III. An Applied Classification Models in Diagnosing Breast Cancer Grades

In data mining, classification is one of the most important tasks. It maps the data in to predefined targets. It is a supervised learning as targets are predefined. The aim of the classification is to build a classifier based on some cases with some attributes to describe the objects or one attribute to describe the group of the objects. Then, the classifier is used to predict the group attributes of new cases from the domain based on the values of other attributes. Different methods serve different purposes, each method offering its own advantages and disadvantages. However, most data mining methods commonly used for assign patients to either a "benign" group that is non- cancerous [G0] or a "malignant" group that is cancerous [G1, G2, G3].

In this paper, Applying a CRISP-DM steps in classifying breast cancer data grades using different mining techniques.

Phase One. Business Understanding:

- **First, Defining Problem Statement:** This phase focusing on understanding the system objectives and requirement by converting a knowledge into data mining problem definition: "Creating a Classification Model to Diagnosing a breast Cancer Grades in Egypt".
- **Second, Determining Variables for Diagnosing:** in this research, pathology report variables will use [Clump thickness, Uniformity of cell size, Uniformity of cell shape, Marginal adhesion,].
- **Finally, Determining Tumor Grades:** (In Egypt) Breast Cancer Grades consists of [G0, Benign] and [(G1, G2, G3) , Malignant].

Table 1: Attributes Description

	Attribute	Description
1	Id	It represent a patient id
2	Age	It represent a patient age
3	Location	It represent a patient job
4	Occupation	It represent a patient job
5	Clump Thickness	Benign cells tend to be grouped in monolayers, while cancerous cells are often grouped in multilayers.
6	Uniformity Of Cell Size	Cancer cells tend to vary in size. That is why these parameters are valuable in determining whether the cells are cancerous or not.
7	Uniformity Of Cell Shape	Cancer cells tend to vary in shape. That is why these parameters are valuable in determining whether the cells are cancerous or not.
8	Marginal Adhesion	Normal cells tend to stick together. Cancer cells tends to loose this ability. So loss of adhesion is a sign of malignancy in other words if the glands is in clear shape so it tends to be benign , and if it is faded glands so it tends to be malignant.
9	Single Epithelial Cell Size	Is related to the uniformity mentioned above. Epithelial cells that are significantly enlarged may be a malignant cell.
10	Bland Chromatin	Describes a uniform "texture" of the nucleus seen in benign cells. In cancer cells the chromatin tends to be more coarse.
11	Normal Nucleoli	Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become more prominent, and sometimes there are more of them.
12	Bare Nuclei	This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors. if an atom is deprived of all its extra nuclear electrons orbiting around the nucleus then it is referred as bare nucleus.
13	Mitosis	Pathologist examines the sample 10 times in 10 different places to determine if there is a cell division or not.
14	Grade Class	It's a target attribute which consists of 4 groups of grade : where G0 mean that there is no malignant cells (benign class) , G1 mean that patient have a first grade of malignant class , G2 mean that patient have a second grade of malignant class, G3 mean that patient have a third grade of malignant class.

Phase Two. Data Understanding:

- **Data Collection:** collect data from literature review from previous benchmark breast cancer datasets, Collecting a Pathology Reports for patients in NCI (Egypt), making interviews within Pathologists, Collect Statistics and books from NCI.
- **Loading Data:** using a Microsoft Excel sheets to enter data.
- **Exploring Data:** this step is very important for data to know the correlation between variable. After applying a correlation matrix, as showed in Table 2.

Table 2: Correlation Matrix of Breast Cancer Dataset, NCI

	<u>Clump</u>	<u>Ucellsize</u>	<u>ucellshap</u>	<u>mgadhesi</u>	<u>sepics</u>	<u>bnuclei</u>	<u>bchrom</u>	<u>normnu</u>	<u>mitoses</u>	<u>Grade Class</u>
Clump	1.0									
Ucellsize	0.61183	1.0								
Ucellshap	0.61077	0.88446	1.0							
Mgadhese	0.45021	0.67493	0.64324	1.0						
Sepics	0.47051	0.71890	0.69602	0.57785	1.0					
Bnuclei	0.55306	0.66628	0.68761	0.62418	0.56588	1.0				
Bchrom	0.53701	0.73729	0.70701	0.65584	0.62110	0.68078	1.0			
Normnu	0.49444	0.67272	0.68777	0.57245	0.62238	0.54901	0.64622	1.0		
Mitoses	0.25084	0.35193	0.33955	0.35446	0.38108	0.24891	0.30257	0.33891	1.0	
Grade Class	0.49733	0.66479	0.65705	0.56915	0.58172	0.57196	0.61098	0.57299	0.32103	1.0

From the previous table, Results show that Mitoses have a low correlation percentage between it and other variables.

Phase Three. Data Preparation:

- **Data Preprocessing:** this phase includes a removal of duplicate records.
- **Handling Missing Values:** by removal of missing data and removing record for missing data in more than 3 cells.
- **Data Cleaning and Transformation:**
 - **Numerical Transformation and Grouping:** grouping some attributes values such as uniformity of cell size values to $[x < 2, 2 < x < 5, x > 5]$ and converting some attributes values to numerical values such as Bare nuclei values [small, moderate, prominent large] to $[1, 2, 3]$ reduce the data complexity while analysis data.

Table 3: Attributes Values and Scoring

Attribute	Values	Scoring
Clump Thickness	Little , Moderate , More	1 , 2 , 3
Uniformity Of Cell Size	$x < 2, 2 < x < 5, x > 5$	1 , 2 , 3
Uniformity Of Cell Shape	Regular or Normal, Medium regularity, Irregular or Abnormal	1 , 2 , 3
Marginal Adhesion	$x > 70 \%, 10 \% < x < 70 \%, x < 10 \%$	1 , 2 , 3
Single Epithelial Cell Size	Small , moderate , Large	1 , 2 , 3

Bland Chromatin	Mild , Moderate, Coarse	1 , 2 , 3
Normal Nucleoli	No visible nucleoli , Medium visibility, Visible nucleoli (prominent)	1 , 2 , 3
Bare Nuclei	Small, moderate, prominent large	1 , 2 , 3
Mitosis	No.of.mit. < 10, 10 < No.of.mit. < 20, No.of.mit. > 20	1 , 2 , 3
Grade Class	G0 , G1 , G2 , G3	0,1 ,2 ,3

- **Description of Data:** the dataset contain 699 records of patients in Egypt from National Cancer Institute, Cairo. Each record represents patient record, represented by a patient id, age, location and lists pathologist parameters.

- **Data Pre-Analysis (Data Statistics):**

Breast cancer dataset represented many grades for beginn [G0] and malignant [G1, G2, G3] groups, figure 3 show that

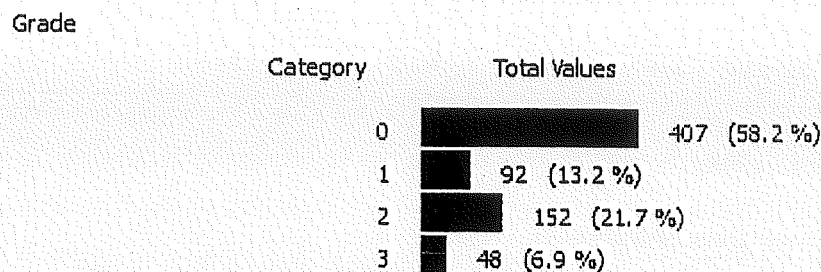


Figure 3: Chart of Benign and Malignant Cases in NCI Data

From the previous figure , Benign Class [G0] represented 407 case (58.2%) and Malignant Class [G1, G2, G3] represented [92, 152, 48] case [13.2%, 21.7%, 6.9%].

In this phase, dataset must be divided into two parts for training and validation phase. **Training dataset** represented 66% [460 record] and **validation dataset** represented 34% [239 record].

- **Feature Selection:** Remove some attributes from classification process because pathologist doesn't need those attributes when determining a breast cancer grade [id, age, location, occupation].

Phase Four. Data Modeling:

First, Determining Mining and Statistical Tool [DTREG]: It builds classification and regression decision trees, neural networks, support vector machine (SVM), GMDH polynomial networks, gene expression programs, K-Means clustering, discriminant analysis and logistic regression models that describe data relationships and can be used to predict values for future observations.

Second, Applying Classification Models: In this phase, the classification model is done by using four data mining models [Decision Tree, Multilayer Perceptron, Support Vector Machine, and Gene Expression Programming].

I. Decision Tree

A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item . Choice of a certain branch depends upon the outcome of the test. To classify a particular data item, we start at the root node and follow the assertions down until we reach a terminal node (or leaf). A decision is made when a terminal node is approached. Decision trees can also be interpreted as a special form of a rule set, characterize by their hierarchical organization of rules [14]. The Decision Tree algorithm produces accurate and interpretable models with relatively little user intervention. It can be used for both binary and multiclass classification problems.

II. Multilayer Perceptron

Neural networks (NN) are those systems modeled based on the human brain working. As the human brain consists of millions of neurons that are interconnected by synapses, a neural network is a set of connected input/output units in which each connection has a weight associated with it. The network learns in the learning phase by adjusting the weights so as to be able to predict the correct class label of the input [15]. Neural Network have some benefits Nonlinearly, Fault Tolerant, Uniformity of Analysis and Design and Adaptively. It can be used for both binary and multiclass classification problems.

III. Gene Expression Programming

Genetic algorithms and evolutionary programming are algorithmic optimization strategies that are inspired by the principles observed in natural evolution. Of a collection of potential problem solutions that compete with each other, the best solutions are selected and combined with each other. In doing so, one expects that the overall goodness of the solution set will become better and better, similar to the process of evolution of a population of organisms. Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism [16]. The classifiers obtained through GP are usually understandable and transparent. It can be used for both binary and multiclass classification problems.

IV. Support Vector Machine

Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant feature [17]. SVM classification is based on the concept of decision planes that define decision boundaries. A decision plane is one that

separates between a set of objects having different class memberships. SVM finds the vectors ("support vectors") that define the separators giving the widest separation of classes. SVM classification supports both binary and multiclass targets.[18] [19]

Phase Five: Data Evaluation:

• Models Validation and Evaluation:

To measure a performance for four classification models by two measures: Variance Analysis such as [CV, NMSE, RMSE, MSE, MAE] and analysis time. Where The best model that exceeds the lowest variance error and lowest analysis time in the validation phase.

• Variance Measures

• First: Training Phase:

In this step using many variance analysis measures for training dataset, table 4 and figure 4 represent that.

Table 4: Variance Analysis (Training Data) for NCI dataset

Variance Analysis (Training Data)	DS-TREE	MLP-NN	GEP	SVM
Coefficient of variation (CV)	0.302705	0.308279	0.356564	0.308807
Normalized mean square error (NMSE)	1	0.055462	0.074196	0.055652
RMSE (Root mean squared error)	0.2338497	0.2381551	0.2754569	0.238563
MSE (Mean Squared Error)	0.0546857	0.0567179	0.0758765	0.0569123
MAE (Mean Absolute Error)	0.1036489	0.1644542	0.172876	0.2003422

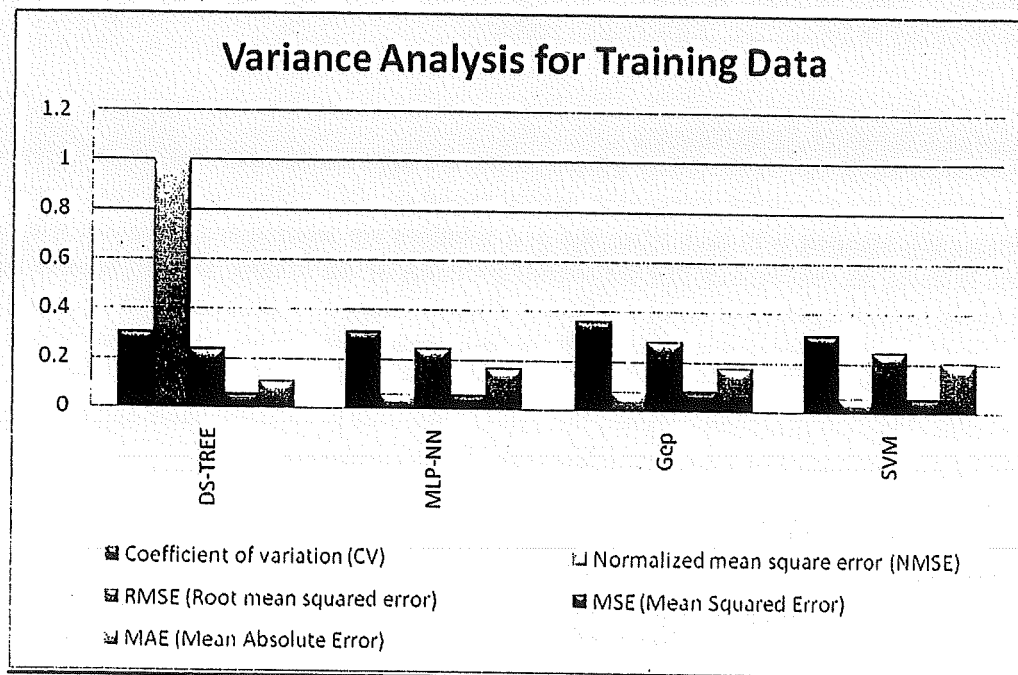


Figure 4: Variance Analysis (Training Data) for NCI dataset

From previous table and chart , decision tree model have lowest error percentage in [CV, RMSE, MSE, MAE] and linear regression model in NMSE, but gene expression programming model have highest error percentage [CV, RMSE, MSE, MAE] and decision tree in NMSE.

• **Second: Validation Phase:**

In this step using many variance analysis measures for validation dataset, table 4 and figure 4 represent that.

Table 5: Variance Analysis (Validation Data) for NCI dataset

Variance Analysis (Validation Data)	DS-TREE	MLP-NN	GEP	SVM
Coefficient of variation (CV)	0.183472	0.317712	0.434972	0.334579
Normalized mean square error (NMSE)	0.038714	0.058908	0.110415	0.065328
RMSE (Root mean squared error)	0.2054714	0.2454426	0.3360298	0.2584727
MSE (Mean Squared Error)	0.0422185	0.0602421	0.112916	0.0668081
MAE (Mean Absolute Error)	0.1074482	0.1696992	0.2197212	0.2102545

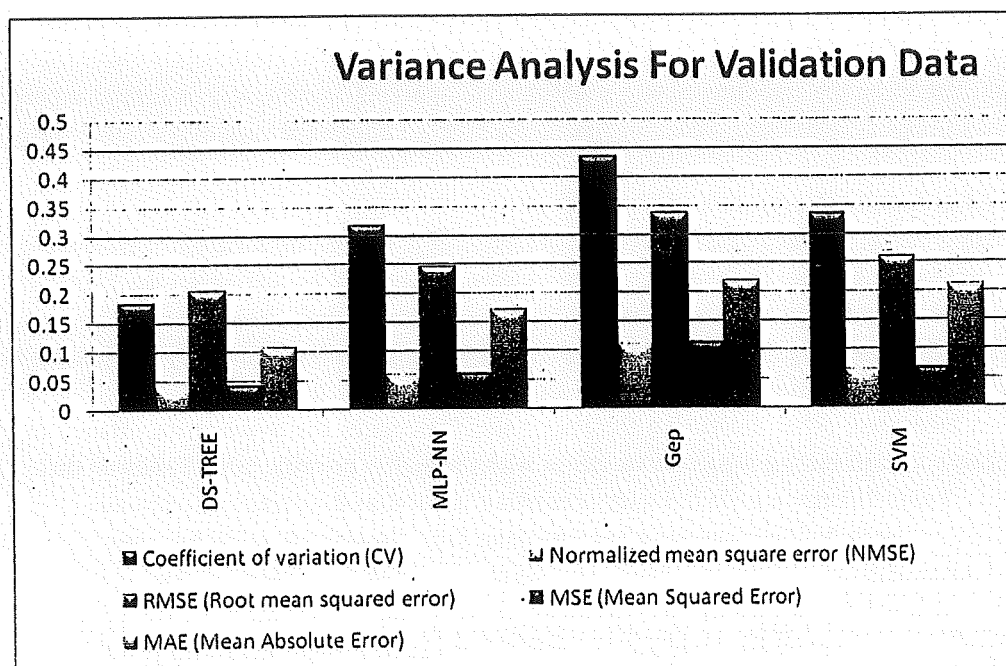


Figure 5: Variance Analysis (Validation Data) for NCI dataset

From previous table and chart, decision tree model have lowest error percentage in [CV, NMSE, RMSE, MSE, MAE], but gene expression programming model have highest error percentage [CV, NMSE, RMSE, MSE, MAE].

• **Analysis Time:**

Analysis time represent a critical performance measure when evaluating or comparing between mining classification techniques because time consuming considered a big problem in any system, table 6 showed which of the four applied techniques that achieve the best or worst analysis time.

Table 6: Analysis Time for four techniques on NCI Data

Analysis Time [minutes : seconds]	DS-TREE	MLP-NN	GEP	SVM
Analysis Time	[00:23]	[01:09]	[01:49]	[02:10]

From previous table Decision Tree have a lowest analysis time at 23 seconds, but support vector machine have a highest analysis time at 2 minutes and 10 seconds.

Phase Six: Deployment:

• **Research Results:**

From the previous analysis for training and validation data using four classification techniques, decision tree model has the best results which have a lowest variance analysis and lowest analysis time where values of [cv, nmse, mse, rmse, and mae] are zero and a lowest analysis time 23 seconds in Validation Phase.

Finally, determine the importance of variables after applying mining and statistical models to diagnosis Breast Cancer data. Then, determine the highly importance variables in all models.

• **Determining Variable Importance:**

Calculating the importance of each predictor variable is measured by different ways depends on the applied technique. In decision tree, calculates importance by adding up the improvement in classification gained by each split that used the predictor. While in multilayer perceptron, gene expression programming and support vector machine calculates importance by using sensitivity analysis where the values of each variable are randomized and the effect on the quality of the model is measured. After measuring the importance calculates the average importance overall four techniques and sorting predicted variables with the highly priority, table 7 showed that.

Table 7: Importance Priority according to total and average importance of each NCI parameter after applying four classification Techniques

VARIABLE IMPORTANCE	DS	MLP	GEP	SVM	Total Import.	Avg. %	Import. Priority
bnuclei	1.99%	100.00%	58.20%	75.94%	236.13%	59.03	3
ucellsize	100.00%	73.70%		100.00%	273.7%	68.43	1
ucellshape	3.38%	55.77%	100.00%	96.44%	207.98%	63.9	2
Mgadhesion	3.78%	47.13%	64.45%	62.62%	225.36%	52	4
Normnucl	3.70%	45.89%	53.25%	65.67%	168.51%	42.13	6
Clump	0.67%	42.98%	4.19%	49.40%	49.4%	12.35	8
Bchromatin	6.29%	37.27%	52.87%	79.14%	175.57%	43.89	5
Sepics		29.86%		43.76%	73.63%	18.41	7
Mitoses	0.32%	7.38%	7.90%	12.56%	28.16%	7.04	9

From the previous table Results showed that the highly importance variables that properly used to diagnosing or classifying breast cancer grades in Egypt are: [ucellsize 68.43%, ucellshape 63.9%, bnuclei 59.03%, mgadhesion 52%, bchromatin 43.89%, Normnucl 42.13%] . On the other side, there are some variables that have smallest importance variables [sepics 18.41%, clump 12.35%, and mitoses 7.04%] which are poorly in classifying our data so prefer ignoring them.

IV. Conclusion and Future Works:

This paper apply a CRISP-DM model steps that widely used in medical mining and also explores the data mining and statistical techniques that help pathologists in diagnosing tumor grades using data mining tool [DTREG]. DS, MLP, GEP and SVM are applied in this paper each of them have each features and differences. Variance metrics like [cv, mae, mse, rmse, and nmse] and analysis time used to evaluate between the four applied techniques. DS classify breast cancer grades dataset with lowest variance metrics and lowest analysis time in validation phase. Those parameters [ucellsize, ucellshape, bnuclei, mgadhesion, bchromatin, and normnuclei] have a highly importance priority to classify breast cancer grades.

Future work will deal with applying a Hybrid data mining techniques models that can be combined Decision tree with Genetic Algorithm or Decision tree with Multilayer Perceptron or Decision tree with Support Vector machine, or mixing three classification techniques with each other and also, we can developing a risk assessment system for breast cancer dataset using mining models.

References

- [1]World Health Organization (2007), cancer control knowledge into action WHO guide for effective programmes Early detection, ISBN: 9241547338
- [2]Bernard W S and Paul K.(2003), World Cancer Report, World Health Organization International Agency for Research on Cancer.
- [3]Ferlay J, Bray F, Pisani P, Parkin DM. GLOBOCAN (2002), Cancer Incidence, Mortality and Prevalence Worldwide IARC CancerBase No 5, version 2.0. Lyon, France: IARC Press; 2004.
- [4]El-Bolkainy M.N.(2000), Topographic pathology of cancer. 2nd edition, NCI, Cairo University.
- [5]Amal S. Ibrahim, Nabil N. Mikhail, Hossam Darwesh, Tarek Heikel (2009), Egypt National Cancer Registry Damietta Profile – 2009 , published july 2011. P. 63-66.
- [6]Amal S. Ibrahim, Nabil N.H. Mikhail, Hussein Khaled, Hoda Baraka, Mohammed Abdeen, Amir Bishara, Sayed Abdel Wahed, Ahmed Abdel Lateef (2008), Cancer Profile in Aswan, Egypt Methodology and Results Chart Book . published April 2010. P. 68 – 77.
- [7]http://en.wikipedia.org/wiki/Breast_cancer
- [8]<http://www.news-medical.net/health/Breast-Cancer-Classification.aspx>
- [9]Osmar R. Zaiane (1999), What are Data Mining and Knowledge Discovery? . University of Alberta.

- [10]The Gartner Group, www.gartner.com.
- [11]David Hand, Heikki Mannila, and Padhraic Smyth (2001), Principles of Data Mining, MIT Press, Cambridge, MA.
- [12]Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi (1998), Discovering Data Mining: From Concept to Implementation, Prentice Hall, Upper Saddle River, NJ.
- [13]<http://www.zentut.com/data-mining/data-mining-processes/>
- [14]Swati Soni , Sini Sibbu (2012), Time Series Analysis of Temporal Data by Classification using Mean Absolute Error , International Journal of Electronics Communication and Computer Technology (IJECCCT), ISSN:2249-7838, Volume 2 Issue 5, September 2012, p.231 – 233.
- [15]Manish Joshi (2012), Classification, Clustering And Intrusion Detection System, International Journal of Engineering Research and Applications (IJERA), ISSN: 2248-9622, Vol. 2, Issue 2, p.961-964
- [16]Michael Goebel, Le Gruenwald (1999), a survey of data mining and knowledge discovery software tools, SIGKDD Explorations journal.
- [17]Päivi Onkamo, Hannu Toivonen (2006), A survey of data mining methods for linkage disequilibrium mapping, National Center for Biotechnology Information.
- [18]A.Shameem Fathima ,D.Manimegalai and Nisar Hundewal (2011), A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue, IJCSI International Journal of Computer Science, ISSN: 1694-0814, Vol. 8, Issue 6., p.322-328.
- [19]Mrs. B. Sujath, Mrs.C.Akil (2012), prediction of breast cancer using data mining, IJCSMR International Journal of Computer Science and Management Research, Vol 1, Issue 3, p.384-389.

Improving the Behavior of QPSO using various

Contraction Coefficient Forms

Tarek M. Elbarbary^{1,*} , Hesham A. Hefny² , Atef M. Abel Moneim³

Abstract

This paper discusses two new alternatives and modified versions of quantum particle swarm optimization as promising solvers for different problems. The first model called "Modified QPSO", which is based on characteristics of QPSO and depends essentially on four different new formulas for the decaying behavior of β contraction expansion coefficient (CEC) with the number of iteration to generate a new solution vector in the search space. This paper summarizes the error comparison results of the proposed MQPSO model with the classical QPSO one. The average error performance improvement is found to be 96.67 %. The second model, called "Weighted mean QPSO", i.e. WMQPSO depends essentially on the first proposed model MQPSO with replacing the mean best position in each iteration by the weighted mean best position. Such an approach causes a great improvement in the error performance over all eight benchmark functions as given in the conclusion. It is found that the average error improved performance of WMQPSO over MQPSO over the eight benchmark function is 87.6647 %. The results of this experiment have shown the superiority of WMQPSO versus other algorithms.

Keywords: *Constrained non linear functions, Contraction Expansion Coefficient, New approach of QPSO, Pattern Search, beta oscillation*

1- Introduction

Swarm Intelligence (SI) is a searching methodology, which is inspired from the social behaviors of natural organisms living in groups or colonies. Particle Swarm Optimization (PSO), method, originally introduced by Eberheart and Kennedy in 1995 [1], is one of the population-based searching approaches. PSO inspired by the natural habits of bird flocking and fish schooling is based on a searching process point of view. The position vector of each particle (bird) represents a potential solution in a D-dimensional search space. During the searching process, particles update their positions and velocities according to the fitness of their historical positions (i.e. cognitive behavior) as well as the population best position (i.e. social behavior). Despite the considerable success of the philosophy of PSO in solving many optimization problems, it suffers a convergence difficulty that makes it hard to find near optimal solutions for several optimization problems [2].

¹ Department of Computer Science, Faculty of Business and Information System, MUST University, Calro, Egypt

² Department of Computer Science ,Institute of Statistical Studies and Research, Cairo University , Glza, Egypt

³ Department of Mathematical Statistics, Institute of Statistical Studies and Research ,Calro University ,Glza ,Egypt

Quantum behaved particle swarm optimization (QPSO) was introduced in 2004 as a new efficient version of PSO. QPSO is inspired by the quantum behavior of atomic particles. In quantum physics, the state of a particle with momentum and energy is determined by its wave functions rather than its position and velocity. By adopting such a philosophy in the position updating equations of the particles in the search space, the convergence process towards best global positions is greatly accelerated compared with the classical PSO [2]. To develop a new approach of Quantum behaved particle swarm optimization that can solve a class of engineering optimization problems while be competitive in both time and efficiency compared with other solution algorithms while, this developed approach will:

- ☐ Emulate in some part the natural phenomena.
- ☐ Be able to solve problems with either continuous or discrete variables.
- ☐ Be able to solve constrained problems and/or unconstrained ones.

Methodology

- ☐ Selecting a number of benchmark problems in engineering optimization area
- ☐ Studying the different algorithms already have been applied to solve these problems
- ☐ Build a new QPSO approach (MQPSO – WMQPSO)
- ☐ Code the problem and build a program using a suitable programming language.
- ☐ Test the program, define the values of the working parameters and evaluate the performance of the algorithm in solving the selected optimization problems.

2. The Proposed Approach for Quantum Particle Swarm MQPSO

We present a new and an alternative quantum behaved particle swarm optimization approach “MQPSO” for solving constrained and unconstrained global optimization problems. This modification is shown to enhance the local search ability of the basic QPSO and to increase the convergence rate. We presents a comparative study between traditional particle swarm optimization PSO, quantum behaved particle swarm optimization (QPSO) and novel model of QPSO named MQPSO for solving global optimization problems. The MQPSO algorithm is based on the characteristics of QPSO, and uses interpolation based recombination operator for generating a new solution vector in the search space by using two modifications:

1. Improving the behavior of classical QPSO by introducing four formulas for adaptively adjusting the contraction expansion coefficient parameter β during the searching process.
2. Proposing a new improved version of QPSO called “Weighted Mean Quantum PSO”.

2.1. Improvement by Contraction Expansion Coefficient

In QPSO algorithm, although the search space of an individual particle at each iteration is the whole feasible solution space of the problem, the loss of diversity of the population is also inevitable due to the collectiveness. This is because according to position equation the search area for each particle is a circle area with centre ‘p’, and $(L/2) * \ln(1/u)$ the radius. Theoretically if ‘u’ is near

0, the radius will be infinite, that is the particle can search in a wider area. However, the probability is very small for the particle to search a far away area from the centre 'p' since the probability for 'u' to get values near 0 is very small. Therefore, the parameter $L=2\beta \cdot |p-x|$ is a very important factor to determine the range of the search area. Introduced in the following is a modification for the parameter. In traditional QPSO, β usually decreases linearly from $\beta_{\max}$ to $\beta_{\min}$ according to generation, so at a later period with β becoming smaller, all particles will exploit near their centers. To achieve trade-off between exploration and exploitation, β is set to vary adaptively in response to the objective values of the particles as follows:

a. Linear Decaying

$$\beta_1 = \frac{(\text{Max Iteration} - \text{Iteration})}{2 \text{ Max Iteration}} + \frac{1}{2} \quad (1)$$

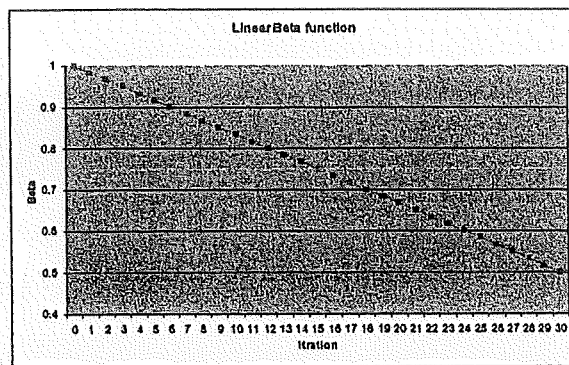


Fig 1 : Linear Beta Coefficient

b. Damping Oscillatory Decaying

$$\beta_2 = \left(\frac{F}{\left(\text{Iteration} + \frac{\text{Beta wave length}}{2} \right)} \right)^{\frac{1}{5}} \times \left(\cos \left(\frac{\text{Iteration number}}{\pi \text{ Beta wave length}} \right) + 1 \right) + \frac{1}{3} \quad (2)$$

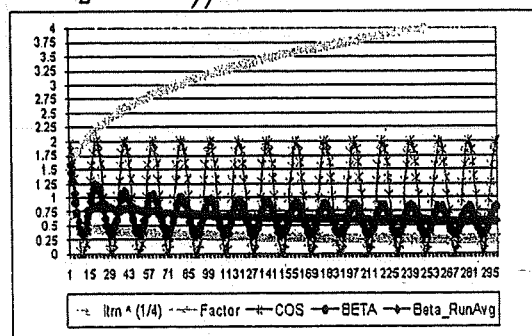


Fig 2 Damping Beta Coefficient

c. Exponential Oscillatory Decaying

$$\beta_3 = \left(1 + \cos \left(\frac{\text{Iteration}}{\pi \text{ Beta wave length}} \right) \right) * e^{-\left(\frac{\text{Iteration}}{\text{Beta Wave Length}} \right)} + \frac{1}{3} \quad (3)$$

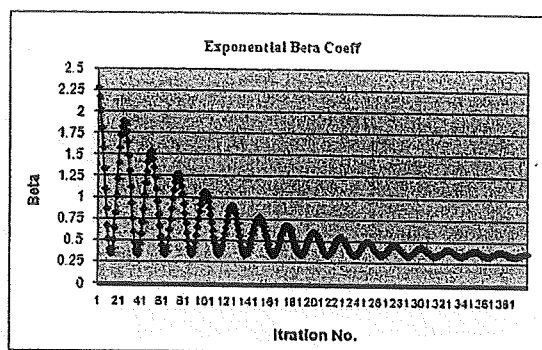


Fig 3 Exponential Beta coefficient

d. Saw Tooth Beta

$$\beta_4 = \frac{1}{2} \left(\frac{\text{Beta Wave length} - \left(\frac{\text{Iteration}}{\text{Beta Wave Length}} \right)}{\text{Beta Wave Length}} \right) + \frac{1}{2} \quad (4)$$

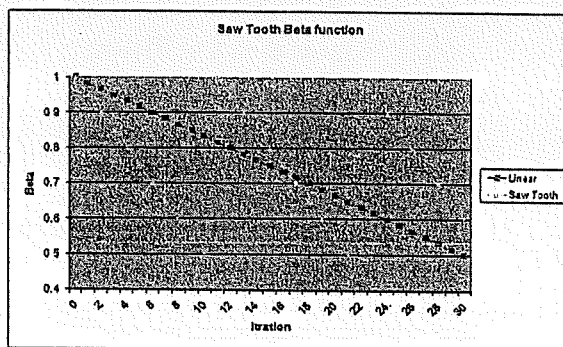


Fig 4 Saw Tooth Beta coefficient

We presents different forms of beta oscillation to decaying the value of beta to get different performance of modified quantum behaved particle swarm optimization, the next tables represent the performance of the new model of QPSO which is MQPSO with different Algorithms regardless of population size. From figure 5, it is clear that the proposed MQPSO algorithm is superior to all other competitor algorithms in terms of average values and standard deviation.

Table 1 (a): Mean best fitness values and standard deviations by algorithms and problems

Algorithm	Sphere	Rosenbrock	Schwefel	Rastrigin
PSO [10]	2.68E-11	49.065	-8514.355	92.551
(std. dev.)	1.60E-10	37.1	447.179	21.626
Gaussian PSO[10]	2.50E-04	36.352	-7478.383	125.335
(std. dev.)	0.002	35.486	566.152	28.576
Gaussian bare[10]	2.44E-41	33.034	-8503.117	79.258
(std. dev.)	1.20E-40	30.145	495.83	1.276
Levy PSO[10]	1.09E-11	43.626	-9478.158	47.56
(std. dev.)	9.93E-11	37.9	421.141	16.984
QPSO[10]	4.61E-23	51.004	-9285.4	25.415
(std. dev.)	2.49E-22	49.596	796.491	6.869
MQPSO	3.85E-56	7.27935449	7.18E-05	2.003142
(std. dev.)	3.52E-54	5.2654	0.02233	0.8258

Table 1 (b) : Mean best fitness values and standard deviations by algorithms and problems

Algorithm	Ackley	Griewank	Rotated ellipsoid	different power
PSO [10]	2.372	0.087	0.408	6.27E-04
(std. dev.)	1.918	0.236	52.36	1.26E-02
Gaussian PSO[10]	11.4	0.579	0.398	3.70E-06
(std. dev.)	3.192	1.094	0.002	3.10E-02
Gaussian bare[10]	1.276	0.019	2.2	0.398
(std. dev.)	1.215	0.028	1.293	2.123
Levy PSO[10]	0.398	0.029	0.00326	8.32E-03
(std. dev.)	2.2	0.04	0.04532	0.0065
QPSO[10]	2.78E-12	0.0101	1.06E-02	9.71E-07
(std. dev.)	7.11E-12	0.0122	2.10E-03	0.0614
MQPSO	8.88E-16	9.83E-73	4.66E-04	9.72E-36
(std. dev.)	3.51E-15	8.95E-70	3.20E-03	3.84E-09

The convergent figure suggests that the MQPSO algorithm can reach the true global solution in all problems in a fewer number of generations less than the maximum predetermined number of generations. Therefore, the proposed formulas algorithm is proven to be an effective, powerful approach for solving unconstrained global optimization problems.

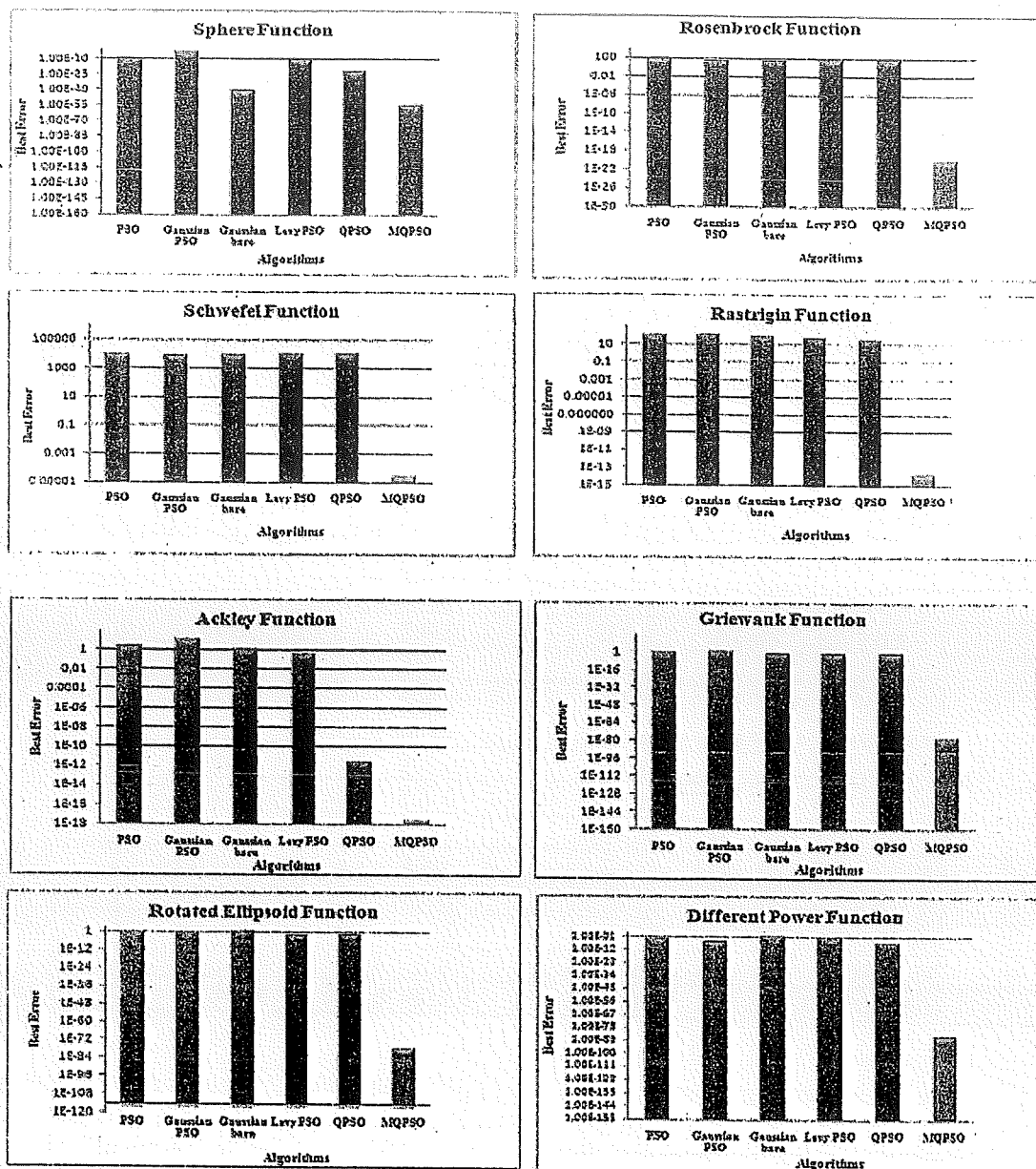


Fig 5: the performance of MQPSO comparing with other programs

2.2. Improvement by Weighted mean

When the lagged particles are chasing after their colleagues, say converging to P_g , the position $mbest$ will be approaching P_g slowly. The distances between position $mbest$ and the personal best positions of a particle near P_g do not decrease quickly, decelerating the convergences of the particles near P_g , and making them explore globally around P_g temporarily until the lagged ones are close to P_g .

$$\text{weighted mean}(x_{t+1}) = \frac{\sum_{i=1}^n \text{fitness value} \times x_t}{\sum_{i=1}^n \text{fitness value}} \quad (5)$$

Table 2 Experimental results of mean best Error values by algorithms and problems

Functions	Algorithms	10	20	30
Sphere	PSO	2.66E-07	8.09E-06	1.32E-04
	QPSO	2.42E-15	1.46E-15	8.60E-19
	MQPSO	9.93E-63	9.96E-63	3.85E-56
	WMQPSO	9.32E-151	4.95E-103	1.42E-65
Rosenbrock	PSO	0.22446	20.34	64.68
	QPSO	0.4924	0.20014	7.73
	MQPSO	1.175E-15	4.845E-10	0.00602469
	WMQPSO	3.25E-26	2.92E-11	1.22E-01
Schwefel	PSO	2290.57	4844.77	7819.25
	QPSO	730.38	240.9	434.59
	MQPSO	2.55E-05	3.83E-03	7.18E-05
	WMQPSO	2.55E-05	2.55E-05	2.55E-05
Rastrigin	PSO	3.24E-06	3.32E-03	6.81
	QPSO	4.97	13.52	19.98
	MQPSO	5.65E-13	3.06E-08	2.03106
	WMQPSO	1.42E-14	2.95E-08	1.9899
Ackley	PSO	1.19	1.77E-03	3.94
	QPSO	2.66E-04	9.48E-08	1.94E-03
	MQPSO	4.74E-15	5.31E-15	4.74E-15
	WMQPSO	7.59E-18	1.28E-17	5.30E-15
Griewank	PSO	6.11E-06	9.92E-08	1.29E-06
	QPSO	6.68E-24	9.74E-14	4.33E-14
	MQPSO	1.40E-62	9.89E-62	6.15E-33
	WMQPSO	9.78E-151	3.13E-100	4.23E-38
Rotated ellipsoid	PSO	1.07E-06	5.60E-03	0.408
	QPSO	1.09E-09	2.18E-09	0.01058
	MQPSO	1.68E-06	1.64E-04	0.000466
	WMQPSO	5.30E-66	5.55E-21	1.73E-03
Sum different power	PSO	4.48E-05	5.13E-04	6.27E-04
	QPSO	1.45E-06	2.79E-08	9.71E-07
	MQPSO	7.62E-11	7.60E-11	1.42E-10
	WMQPSO	9.56E-151	7.08E-124	2.73E-74

Therefore, in the MQPSO approach with weighted mean best position, the particle swarm never abandons any lagged particle and seems to be more intelligent and more cooperative social organism. In a word, the wait mechanism among particles enhances the global search ability of MQPSO greatly

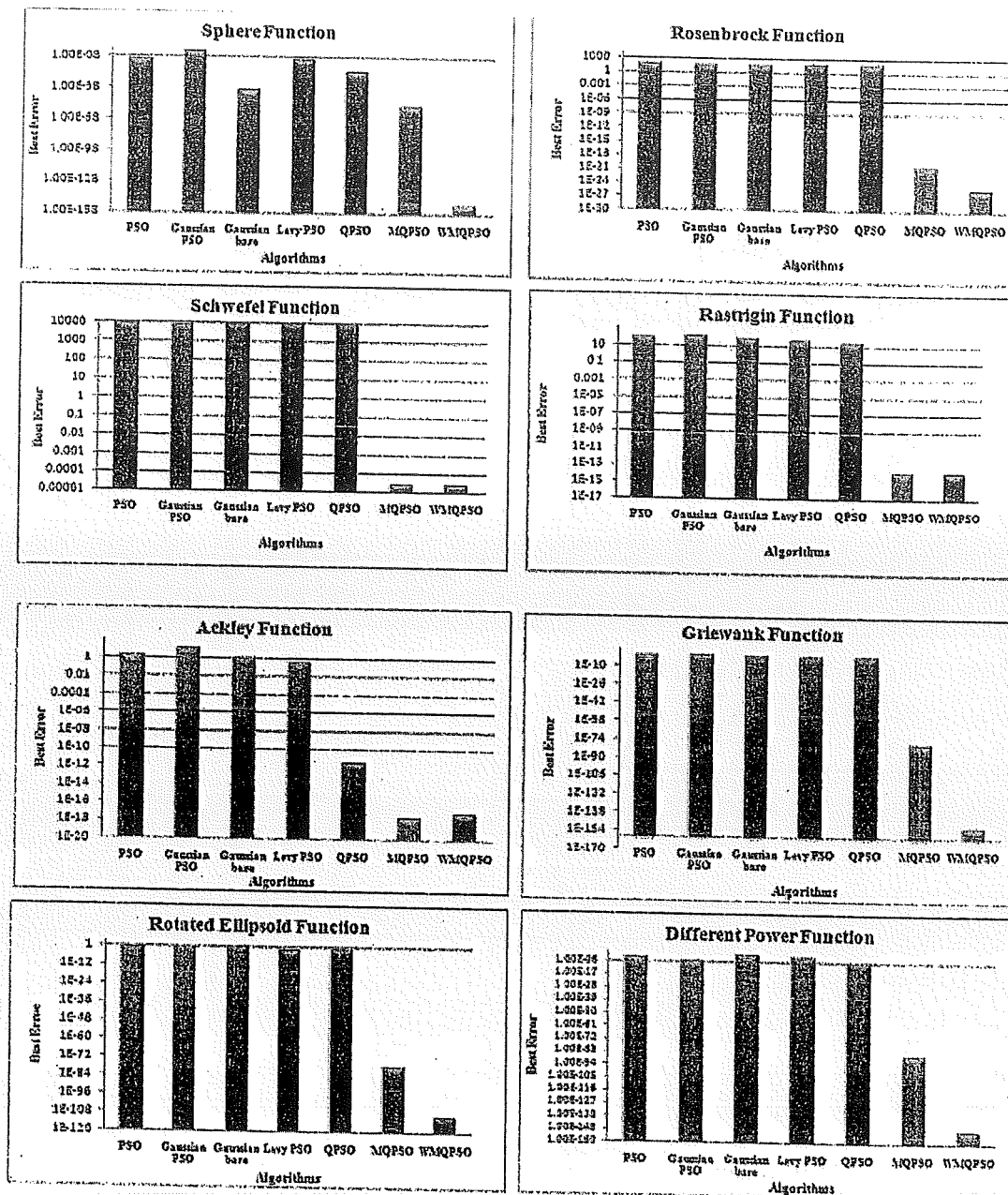


Fig 6: The performance of WMQPSO with other programs

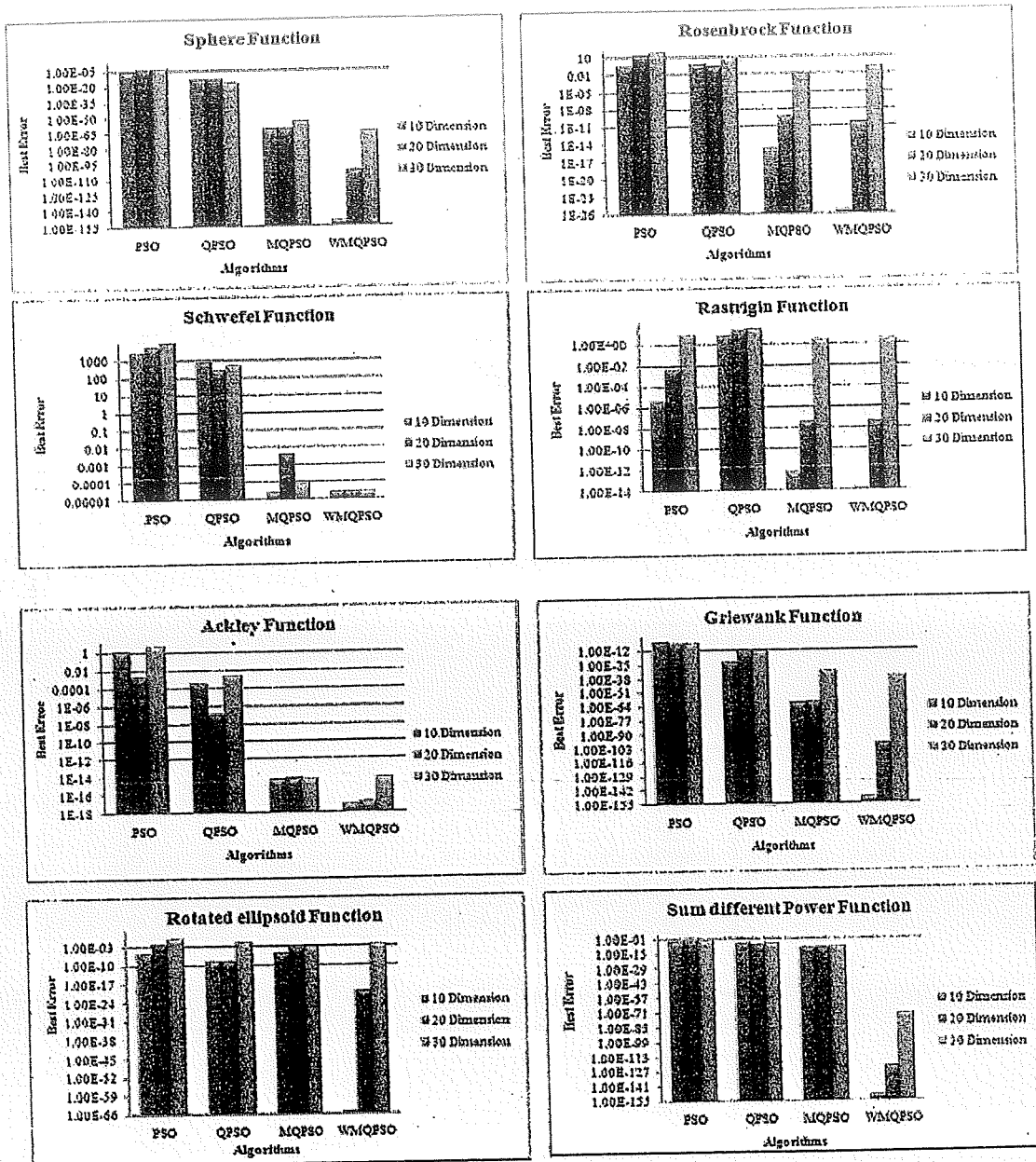


Fig 7: The performance of WMQPSO with other programs with different dimensions

It can be observed from figures 6, 7 and 8, that the MQPSO algorithm costs much less computational effort than the QPSO and PSO algorithms. Therefore, as a lower number of function evaluations cross ponds to a faster convergence, the MQPSO with Saw Tooth beta form algorithm is the fastest one among all competitor algorithms.

Table 3 : The Features of WMQPSO with respect to Beta Forms

Functions	Dim	Features	β_1	β_2	β_3	β_4
Sphere	10	Worst	4.728229046	3.852036715	3.05324173	2.886556149
		Median	0.000251027	0.000272158	0.000408978	0.000310188
		Best	4.21E-39	8.75E-11	1.94E-62	9.05E-65
		Standard Dev	5.96E-37	7.13E-10	6.99E-25	3.52E-22
		MSE	4.21E-39	5.15E-10	1.92E-32	5.06E-25
	20	Worst	2.54E-12	5.53E-13	1.89E-09	5.22E-10
		Median	6.35E-25	6.77E-16	2.95E-22	1.87E-27
		Best	9.96E-63	1.35E-29	1.49E-22	4.06E-58
		Standard Dev	3.85E-48	1.88E-37	6.99E-25	3.52E-22
		MSE	1.25E-43	1.04E-18	5.04E-18	1.55E-37
	30	Worst	6.98E-09	4.37E-09	4.86E-10	8.34E-08
		Median	1.57E-23	7.27E-13	7.94E-12	1.29E-17
		Best	3.85E-56	3.53E-18	6.63E-14	5.86E-24
		Standard Dev	7.18E-49	7.63E-10	7.11E-13	3.52E-22
		MSE	1.25E-43	1.15E-15	1.71E-07	7.32E-16
Rosenbrock	10	Worst	0.631663144	0.500723064	0.493498743	0.493498743
		Median	0.239317968	0.239250079	0.239250079	0.239250079
		Best	5.46E-21	3.12E-12	2.02E-03	5.93E-19
		Standard Dev	3.39E-20	4.93E-07	4.64E-10	8.60E-07
		MSE	1.15E-15	1.70E-07	1.15138	7.31E-16
	20	Worst	43.95019913	43.95796967	44.10442352	21.04797554
		Median	10.36738491	9.916528702	9.529525757	8.945329666
		Best	1.27E-09	2.85E-02	4.28E-02	1.36E-05
		Standard Dev	1.01E-09	6.59E-10	2.79E+00	2.92E-03
		MSE	1.70E-07	0.03542	0.0129	0.1615712
	30	Worst	22.55408478	22.74043465	20.58324051	19.54739761
		Median	11.22849178	10.88460636	11.01218987	10.99787807
		Best	3.28E-01	6.02E-03	1.06E+00	4.55E-01
		Standard Dev	2.64E+00	4.09E-04	6.30E+00	8.99E+00
		MSE	0.88675	0.3278	3.11551	3.4551
Schwefel	10	Worst	833.1967773	790.0297852	779.2312012	778.5339355
		Median	240.9	434.59	304.3432	34.26
		Best	2.95E-04	2.55E-05	2.95E-04	2.24E-04
		Standard Dev	3.42E-02	3.42E-03	3.42E-02	3.42E-03
		MSE	8.52E-02	9.95E-04	5.82E-03	2.67E-02
	20	Worst	343.1135254	335.515625	346.1623535	335.1135254
		Median	113.56	135.6	89.875	97.34752
		Best	3.82E-03	3.82E-03	1.30E+00	2.55E-05
		Standard Dev	2.43E+00	2.301	2.32E-01	3.00E-03
		MSE	1.000342	9.37E-03	10.3564	3.89E-03
	30	Worst	250.2034912	249.4974976	221.8868637	207.4401703
		Median	81.36849594	61.22621918	54.90084839	8.954781532
		Best	3.82E-03	7.18E-05	1.40E+00	7.18E-05
		Standard Dev	2.00E-03	3.45E-03	2.43E+00	3.82E-03
		MSE	0.0452	3.90E-03	2.346	2.43E-02

Rastrigin	10	Worst	3.17233181	2.079618216	1.157348633	2.134899139
		Median	0.000275314	0.000334899	0.00019374	0.00019374
		Best	1.42E-14	1.42E-14	2.25E-04	1.42E-14
		Standard Dev	2.12E-22	2.12E-22	2.65E-11	2.12E-22
		MSE	1.70E-07	1.58E-12	0.02851659	1.91E-12
	20	Worst	3435.113525	94.79501343	87.64680862	82.02971649
		Median	3.979836941	7.94076	5.694076	3.979984076
		Best	2.84E-14	2.83E-14	7.61E-02	8.53E-14
		Standard Dev	2.65E-11	2.51E-15	1.32E-09	2.10E-11
		MSE	1.32E-09	1.03E-12	1.146712	2.65E-10
	30	Worst	140.9372177	114.1296272	94.79501343	82.02971649
		Median	6.99959898	6.977294683	6.964714885	3.979836464
		Best	1.14E-13	6.82E-13	2.00E+00	3.41E-13
		Standard Dev	2.10E-11	2.51E-09	6.53E-02	2.51E-12
		MSE	3.56E-12	3.56E-10	2.385256	3.56E-12
Ackley	10	Worst	6.483581734	5.686917496	4.652520657	3.760482693
		Median	3.41E-12	3.41E-12	2.44E-13	1.51E-14
		Best	2.39E-18	4.12E-18	4.12E-18	4.12E-18
		Standard Dev	3.66E-26	4.66E-14	4.66E-14	4.66E-14
		MSE	3.57E-12	5.10E-17	9.10E-17	5.10E-17
	20	Worst	21.0589838	8.683698734	7.763530254	6.765764872
		Median	1.32E-10	6.13E-11	5.81E-11	2.27E-12
		Best	1.11E-17	9.32E-18	4.43E-11	2.49E-17
		Standard Dev	3.58E-25	7.22E-11	3.01E-06	7.13E-09
		MSE	5.10E-17	9.09E-12	1.73E-18	1.04E-13
	30	Worst	5.709776402	4.964224696	2.978392199	7.42E-06
		Median	1.74E-10	4.45E-14	3.66E-09	4.78E-15
		Best	8.88E-16	4.77E-15	2.14E-10	4.79E-15
		Standard Dev	7.13E-09	4.66E-14	1.89E-06	1.03E-07
		MSE	5.10E-15	7.19E-09	1.08E-07	3.56E-12
Griewank	10	Worst	1.301428318	1.012198806	0.733679354	0.526573718
		Median	4.05E-19	3.95E-23	2.30E-22	2.95E-25
		Best	6.94E-75	7.68E-79	9.49E-65	9.51E-76
		Standard Dev	5.44E-27	1.15E-18	2.73E-28	5.44E-27
		MSE	1.09E-26	2.00E-33	9.29E-31	5.17E-36
	20	Worst	2.825732231	0.000219177	0.000113499	9.16E-05
		Median	2.74E-23	2.54E-23	5.88E-12	4.47E-28
		Best	9.99E-65	3.59E-45	2.92E-25	3.35E-41
		Standard Dev	9.29E-31	2.00E-33	2.73E-28	1.04E-13
		MSE	7.74E-34	2.72E-39	1.13E-22	1.09E-26
	30	Worst	1.902339399	1.452001989	1.075194836	0.96639657
		Median	8.00E-23	3.19E-10	1.80E-04	5.86E-30
		Best	9.37E-38	5.50E-22	3.51E-11	9.83E-73
		Standard Dev	5.44E-27	1.15E-18	7.13E-09	9.90E-62
		MSE	1.09E-26	9.09E-12	2.17E-10	7.74E-34

Rotated Ellipsoid	10	Worst	1.193498731	0.027668798	0.021795567	0.00205927
		Median	0.001013886	0.001021944	7.31395E-05	4.1534E-06
		Best	9.55E-79	1.60E-25	1.26E-16	4.96E-48
		Standard Dev	2.45E-34	5.21E-20	3.51E-19	2.77E-17
		MSE	2.72E-39	1.66E-18	3.80E-05	1.09E-26
	20	Worst	3.312782526	3.312782526	2.000044763	1.971015811
		Median	0.000274034	0.000273978	0.000273791	0.000273791
		Best	3.07E-10	1.44E-06	3.70E-04	3.81E-07
		Standard Dev	3.51E-09	1.89E-08	1.89E-03	1.32E-10
		MSE	6.67E-09	3.80E-05	4.24E-03	0.0001599
	30	Worst	12.77072525	19.41787823	3.538511038	3.519468796
		Median	0.022868059	0.016253443	0.011172566	0.009319704
		Best	1.77E-02	9.52E-04	8.88E-04	4.15E-04
		Standard Dev	5.21E+00	3.51E+00	1.89E+00	3.91E+00
		MSE	3.748009	1.7154	0.8976	2.7154
Different Power	10	Worst	0.498383641	0.108917102	0.108901702	0.120225303
		Median	9.9752E-11	1.83042E-11	7.33932E-10	4.22555E-11
		Best	9.43E-87	9.88E-86	9.88E-18	9.17E-86
		Standard Dev	2.92E-25	1.71E-25	2.80E-14	1.71E-25
		MSE	2.94E-50	1.81E-48	3.69E-17	5.26E-35
	20	Worst	2.630236626	1.361049533	0.001521831	0.000467385
		Median	2.62E-10	2.59E-10	2.35E-10	2.34E-10
		Best	8.29E-76	9.88E-76	3.14E-18	9.44E-76
		Standard Dev	1.14E-28	1.71E-25	9.79E-02	2.92E-25
		MSE	1.24E-37	1.84E-30	1.25E-14	7.24E-35
	30	Worst	2.374967575	1.991910696	0.004001441	0.001376212
		Median	1.25E-09	9.69E-10	8.61E-10	7.13E-10
		Best	1.32E-49	2.94E-50	3.69E-17	1.81E-48
		Standard Dev	1.44E-32	5.26E-35	2.80E-14	1.24E-24
		MSE	2.93E-25	7.87E-28	2.79E-14	1.13E-28

From Table 3 it is clear that the proposed MQPSO algorithm is superior to all other competitor algorithms in terms of average values and standard deviation. Furthermore, the results showed that MQPSO algorithm outperformed basic QPSO algorithm in all functions. The performance of most of the evolutionary algorithms deteriorates with the growth of dimensionality of the search space. As a result, in order to test the performance of the WMQPSO, MQPSO and basic PSO and QPSO algorithms, the scalability study was conducted. The benchmark functions were studied at D = 30 dimensions. The population size was chosen as NP =128 for all dimensions. Convergence Figures present the convergence of proposed MQPSO and WMQPSO algorithms in term of the average fitness values of the best vector found during generations for selected benchmark functions.

Moreover it also outperformed PSO algorithm in all functions. Additionally, the MQPSO algorithm outperformed Gaussian PSO algorithm in all function. It is worth mentioning that the MQPSO algorithm considerably improves the final solution quality and it is extremely robust since it has a small standard deviation on all functions.

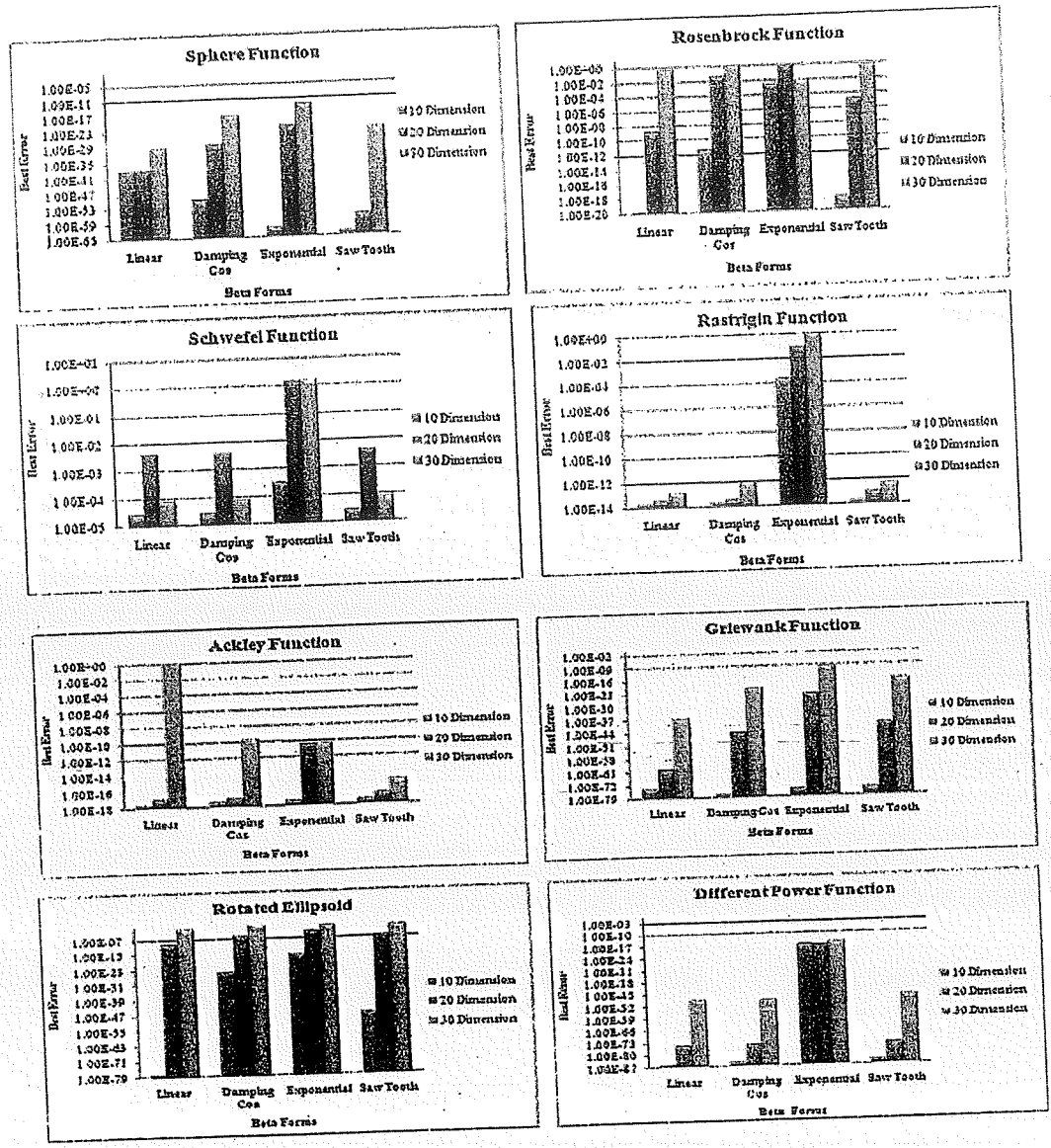


Fig 8: he performances of WMQPSO with different Dimensions

Contrarily the MQPSO algorithm achieved the accuracy level in all 100 runs-with all functions. Moreover, the number of successful runs was also greatest for the MQPSO algorithm over all functions. Thus, this indicates the higher robustness of the proposed algorithm as compared to other algorithms and also proves the capability in maintaining higher diversity with an improved convergence rate. Similarly, consider the convergence characteristics of benchmark functions presented in Figure 6 to figure 8 it is clear that the convergence speed of the modified QPSO with new betas.

Figure 8 demonstrated that WMQPSO still has a better global exploring ability for promising area and a better local exploiting ability for locating the optimal solution regardless the effect of the population size. It can be concluded that as the number of dimensions increases, the performance of the basic PSO and QPSO algorithms rapidly deteriorates whereas the performance of MQPSO and WMQPSO algorithms slightly decreases. The results show that the proposed WMQPSO

algorithm outperformed the basic PSO and QPSO techniques in all function by remarkable difference while it outperformed the MQPSO algorithm in most test functions, for various dimensions. We note that the convergence speed of the modified QPSO with new betas formulas algorithm is fast at the early stage of the optimization process for all functions with different shapes, complexity and dimensions. The convergent figure suggests that the MQPSO algorithm can reach the true global solution in all problems in a fewer number of generations less than the maximum predetermined number of generations.

5. Conclusion

In this paper, two new models for quantum particle swarm optimization are proposed. The first model called "Modified QPSO" which depends essentially on four different new formulas for the decaying behavior of β coefficient with the number of iterations.

Table 4: General Comparison of the performance of the proposed Approached QPSO and the standard QPSO

FUNCTIONS	QPSO [2]	MQPSO	$\frac{E_{QPSO} - E_{MQPSO}}{E_{QPSO}} \times 100$
Sphere	4.61E-23	3.85E-56	100%
Rosenbrock	5.10E+01	7.279354491	85.72%
Schwefel	1.40E+00	3.95-04	99.97%
Rastrigin	2.54E+01	2.003142	92.11%
Ackley	2.78E-12	8.88E-16	99.97%
Griewank	1.01E-02	3.69E-10	100.00%
Rotated ellipsoid	1.06E-02	4.66E-04	95.60%
Different Power	9.71E-07	9.72E-18	100%
Average Error Improvement Performance			96.67 %

Table 5: General Comparison of the performance of MQPSO and WMQPSO

FUNCTIONS	MQPSO	WMQPSO	$\frac{E_{MQPSO} - E_{WMQPSO}}{E_{MQPSO}} \times 100$
Sphere	3.85E-56	1.42E-65	100 %
Rosenbrock	7.279354491	1.22E-01	98.32 %
Schwefel	3.95-04	2.55E-05	93.5 %
Rastrigin	2.003142	1.14E-13	100 %
Ackley	8.88E-16	1.28E-17	98.558%
Griewank	3.69E-10	6.15E-38	100%
Rotated ellipsoid	4.66E-04	4.15E-04	10.94 %
Different Power	9.72E-18	1.81E-48	100%
Average Error Improvement Performance			87.6645 %

Table 4 presents the performance of the proposed QPSO by adaptively adjusting the contraction expansion coefficient parameter β during the searching process along with the standard QPSO. The results of the comparison ensure that the superiority of the proposed algorithms is a near optimum solution. It is found that the average of error performance improved over the 8-benchmark functions for the proposed MQPSO compared with classical QPSO. Table 5 summarizes the error performance

of the two proposed models in this paper MQPSO and WMQPSO over the 8- benchmark functions. It is found that the average error performance improvement of WMQPSO compared with MQPSO is 87.66 %.

6. References

1. J. Kennedy, and R.C. Eberhart. "Particle swarm optimization," in Proceedings of IEEE International Conference on Neural Networks, pp. 1942-8, 1995.
2. J. Sun, B. Feng, and W. Xu. "Particle swarm optimization with particles having quantum behavior," in IEEE Congress on Evolutionary Computation, pp. 325-31, 2004.
3. C. Lin, and P. Feng. "Parameters selection and application of support vector machines based on quantum delta particle swarm optimization algorithm," Automation and Instrumentation, vol. 1, pp. 5-8, 2009.
4. J. Sun, W. Fang, X. Wu, Z. Xie, and W. Xu. "Particle swarm optimization with particles having quantum behavior: analysis of the individual particle and parameter selection," Evolutionary Computation, 2010.
5. L. Liu, J. Sun, D. Zhang, G. Du, J. Chen, and W. Xu. "Culture conditions optimization of hyaluronic acid production by Streptococcus zooepidemicus based on radial basis function neural network and quantum-behaved particle swarm optimization algorithm," Enzyme and Microbial Technology, vol. 44(1), pp. 24-32, 2009
6. M. Clerc. "The swarm and the queen: towards a deterministic and adaptive particle swarm optimization," in Proceedings of the 1999 Congress on Evolutionary Computation, pp. 1957, 2002.
7. M. Donelli, and A. Massa. "Computational approach based on a particle swarm optimizer for microwave imaging of two-dimensional dielectric scatterers," IEEE Transactions on Microwave Theory and Techniques, vol. 53(5), pp. 1761-76, 2005.
8. M. Clerc, and J. Kennedy. "The particle swarm - explosion, stability, and convergence in a multidimensional complex space," IEEE Transactions on Evolutionary Computation, vol. 6(1), pp. 58-73, 2002.
9. Millie Pant, Radha Thangaraj and Ajith Abraham, A New PSO Algorithm with Crossover Operator for Global Optimization Problems, Second International Symposium on Hybrid Artificial Intelligent Systems (HAIS'07), Advances in Softcomputing Series, Springer Verlag, Germany, E. Corchado et al. (Eds.): Innovations in Hybrid Intelligent Systems, Vol. 44, pp. 215-222, 2008.
10. W. Fang, J. Sun, and W. Xu. "Improved quantum-behaved particle swarm optimization algorithm based on differential evolution operator and its application," Journal of System Simulation, vol. 20(24), pp. 6740-4, 2010.
11. X. Zhang, H. Zhang, Y. Zhu, Y. Liu, T. Yang, and T. Zhang. "Using IACO and QPSO to solve spatial clustering with obstacles constraints," in IEEE International Conference on Automation and Logistics, pp. 1699-704, 2009.
12. Y. Shi, and R. Eberhart. "A modified particle swarm optimizer," in The 1998 IEEE International Conference on Evolutionary Computation Proceedings, pp. 69-73, 1998.
13. Y. Shi, and R. Eberhart. "A modified particle swarm optimizer," in The 1998 IEEE International Conference on Evolutionary Computation Proceedings, pp. 69-73, 1999.

Analysis of Advanced Encryption Standard Algorithm with Constant and Dynamic Rotation

*I.A. Ismail, **Galal H. Galal-Edeen, Sherif Khattab and Mohamed Abd Elhamid I. Moustafa El Bahtity

Abstract

Recently, the Rijndael algorithm has been uniform by the National Institute of Standards and Technology (NIST) as the Advanced Encryption Standard (AES). This makes AES a vital and necessary data-protection mechanism for federal agencies in the US and other countries. In AES, rotation occurs in key expansion, ciphering, and deciphering. Rotation is vital for confusion and diffusion, which play an important role in any cryptography technique. Confusion and diffusion make breaking the key complex and difficult. This paper studies the effect of reconfiguring the structure of AES, especially replacing constant rotation with variable rotation. The resulting producing another cipher is called Dynamic Rotation for Advanced Encryption Standard (DRAES). DRAES with variable rotation raises the complexity of the algorithm, and thus, increases the time consumed for brute-force attacks. We measured the diffusion of AES and DRAES algorithms. DRAES reached acceptable level of diffusion faster than AES.

Keywords: AES, DRAES, confusion and diffusion

I. INTRODUCTION

The National Institute of Standards and Technology (NIST), a non-regulatory federal agency, standardized the Advanced Encryption Standard (AES) as Federal Information Processing Standard (FIPS) 197. Prior to AES, the Data Encryption Standard (DES) was the federal standard for block symmetric encryption FIPS 46 in 197 [7]. In June 2003 the US government has approved the use of 128, 192, 256 bit key AES for secret and 192, 256-bit key AES for top-secret information.

Now, after the publication of FIPS 197, AES encryption remains the de facto standard for symmetric encryption, and non-brute-force attacks remain impossible [1, 2], at least for the foreseeable future. To date, most attack methods have focused on weaknesses or characteristics in specific implementations, called side-channel attacks, not on the algorithm itself. However, AES has been remarkably resilient to these attacks [3-6].

* Emeritus dean of faculty of computers and information, Zagazig University

** Faculty of computers and information, Cairo University

In the last ten years, AES has been subject to very intensive cryptanalysis, with best currently known attacks breaking 7, 10, 10 rounds for respective key sizes 128, 192, 256, with very high complexities. In this work, we propose Dynamic Rotation AES (DRAES), a modification and enhancement of the rotation in AES.

The following section contains the evaluation of AES with constant rotation. Dynamic rotation with DRAES is presented in Section III. Diffusion analysis is assessed for both AES and DRAES algorithms in Section IV. Finally, Section V contains conclusions.

II. Evaluation of Advanced Encryption Standard

On the inside of the AES algorithm, processes are executed on a two-dimensional array of bytes called the state. The state consists of four rows of bytes, each containing Nb bytes, where Nb is the block length divided by word size (32 bits). $Nb=4$ for 128-bit block, $Nb=6$ for 192-bit block, $Nb=5$ for 160-bit block, and $Nb=8$ for 256-bit block.

The number of words in the key is called Nk . Ciphering is done by a series of mathematical operations iteratively. The number of rounds (iterations) is represented by Nr , where $Nr=10$ when $Nk=4$, $Nr=12$ when $Nk=6$, and $Nr=14$ when $Nk=8$. In other words, the key length and the number of rounds differ from key size to key size as shown in Table 1. A block size of 128 bits is assumed. The components of the AES encryption algorithm are described next.

Table 1 Common AES variants and their features

Algorithm	Key length (Nk words)	Block Size (Nb words)	Number of rounds (Nr)
Aes-128-bit	4	4	10
Aes-160-bit	5	4	11
Aes-192-bit	6	4	12
Aes-256-bit	8	4	14

1. SubBytes Transform

In the SubBytes phase, the data in the plaintext are substituted by some pre-defined values from a substitution box. The substitution box, which is used commonly, is an AES substitution box (S-box table). Figure 1 demonstrates that the substitution box (S-box) is invertible and non-linear. SubBytes are the only nonlinear operation in AES. Nonlinearity is important for any encryption algorithm.

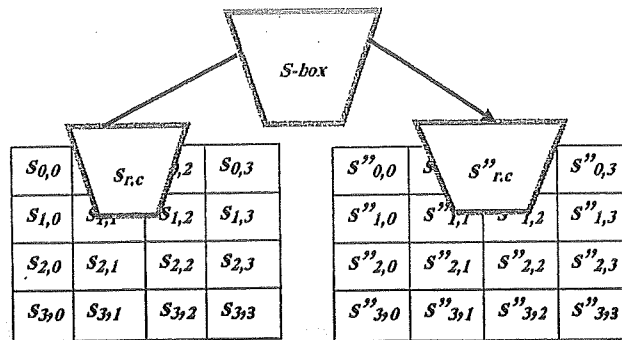


Figure 1 SubBytes () applies the S-box to each byte of the State

2. Shiftrow Transform

In the Shiftrow transformation, the bytes in the last three rows of the State are cyclically shifted with different numbers of offsets (measured in bytes). The first row, Row 0, is not shifted.

Specifically, the ShiftRows transformation proceeds as follows:

$$S'_{r,c} = S_{r, (c + \text{shift}(r, Nb)) \bmod Nb}, \text{ for } 0 < r < 4 \text{ and } 0 \leq c < Nb \quad (1)$$

Where the shift value $\text{shift}(r, Nb)$ depends on row number r , as follows:

$$\text{shift}(1,4) = 1; \text{shift}(2,4) = 2; \text{shift}(3,4) = 3 \quad (2)$$

Figure 2 illustrates the ShiftRows transformation.

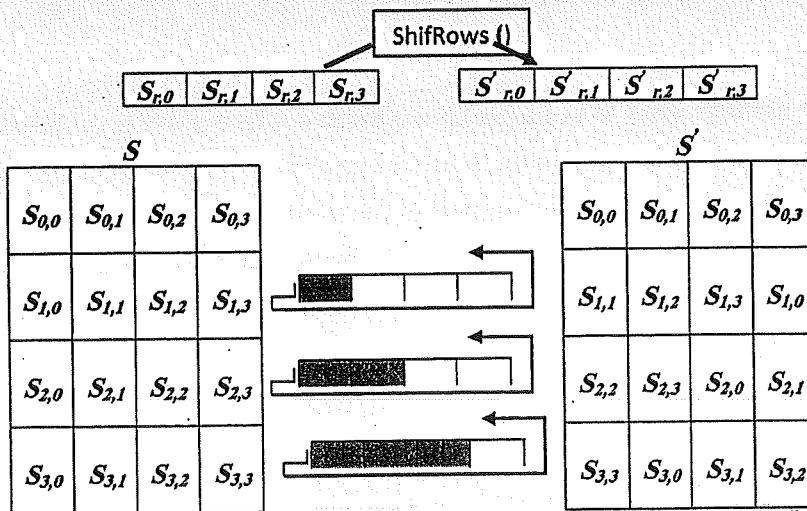


Figure 2 ShiftRows cyclically shifts the last three rows in the State.

3. MixColumns Transform

The MixColumns transformation operates on the State column-by-column, treating each column as a four-term polynomial. The columns are considered as polynomials over $GF(2^8)$ and are multiplied modulo $x^4 + 1$ with a fixed polynomial $c(x)$, given by

$$c(x) = c_0 + c_1x + c_2x^2 + c_3x^3 \quad (3)$$

Where $c_0=0x02$, $c_1=0x01$, $c_2=0x02$, $c_3=0x03$.

This can be written as matrix multiplication: $b(x) = c(x) \otimes a(x)$,

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 02 & 03 & 01 & 01 \\ 01 & 02 & 03 & 01 \\ 01 & 01 & 02 & 03 \\ 03 & 01 & 01 & 02 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

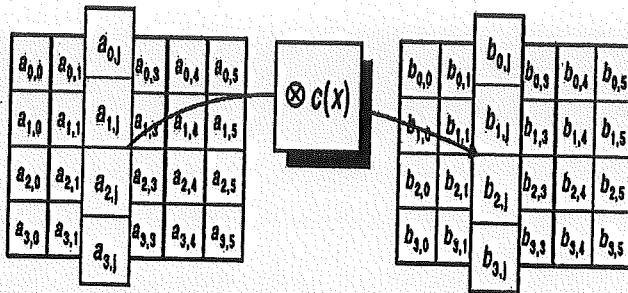


Figure 3 Mixcolumn operates on the columns of the State

4. AddRoundKey Transform

AddRoundKey transformation is as simple as possible and affects every bit of State. The complexity of the round key expansion, plus the complexity of the other stages of AES, ensures security. Each Round Key consists of Nb words from the key schedule. Those Nb words are each added into the columns of the State, such that

$$[S'_{0,c} \ S'_{1,c} \ S'_{2,c} \ S'_{3,c}] = [S_{0,c} \ S_{1,c} \ S_{2,c} \ S_{3,c}] \oplus [W_{round \cdot Nb + c}] \text{ for } 0 \leq c < Nb, \quad (4)$$

Where W_i is a word from the key schedule, and round is a value in the range $0 \leq \text{round} \leq Nr$. In the AES encryption, the initial Round Key addition occurs when $\text{round} = 0$, the application of the AddRoundKey transformation to the Nr rounds of the Cipher occurs when $1 \leq \text{round} \leq Nr$. The process of AddRoundKey transformation is demonstrated in Figure 4, and Figure 5 illustrates the AES encryption and decryption processes.

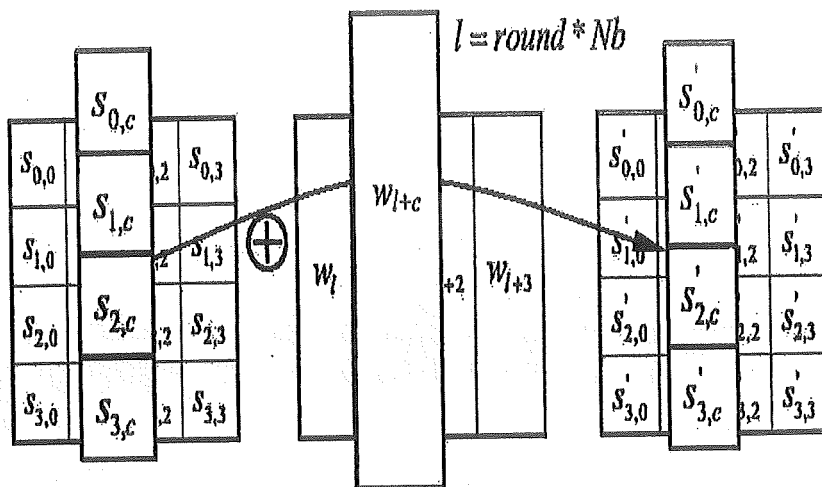


Figure 4 AddRoundKey XORs each column of state with a word from a key schedule.

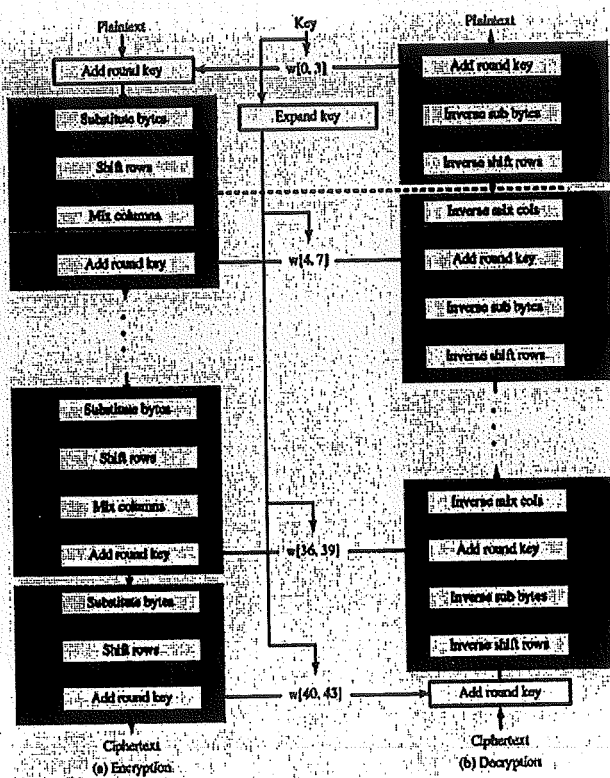


Figure 5 Structure of AES encryption and decryption

5. Key expansion

Key expansion is part of the AES algorithm. It takes as input a 4-word (16-byte) key and produces a linear array of 44 words (176 bytes). This is sufficient to provide a 4-word round key for the initial AddRoundKey stage and each of the 10 rounds of the cipher. The following pseudocode Figure 6 describes key expansion.

Constants: int Nb = 4;

Inputs: int Nk = 4, 6, or 8; // the number of words in the key

array key of 4*Nk bytes or Nk words // input key

Output: array W of Nb*(Nr+1) words or 4*Nb*(Nr+1) bytes // expanded key

Algorithm:

```
void KeyExpansion(byte[] key, word[] W, int Nk) {
    int Nr = Nk + 6;
    W = new byte[4*Nb*(Nr+1)]; int temp; int i = 0;
    while ( i < Nk) {
        W[i] = word (key [4*i], key [4*i+1], key [4*i+2], key [4*i+3]); i++; }
    i = Nk;
    while (i < Nb*(Nr+1)) { temp = W[i-1];
        if (i % Nk == 0) temp = SubWord (RotWord(temp)) ^ Rcon[i/Nk];
        else if (Nk > 6 && (i%Nk) == 4)
            temp = SubWord(temp);
        W[i] = W[i-Nk] ^ temp; i++; } }
```

Figure 6 Implementation of key expansion.

The key is copied into the first four words of the expanded key. The remainder of the expanded key is filled in four words at a time. Each added word W [i] depends on the immediately preceding word, W [i-1], and the word four positions back [i -4]. In three out of four cases, a simple XOR is used. For a word whose position in the w array is a multiple of 4, a more complex function is used. Figure 7 illustrates the generation of the first eight words of the expanded key, using the symbol g to represent that complex function. The function g consists of the following subfunctions:

- Rotation executes a one-byte circular left shift on a word. This means that an input word [b₀, b₁, b₂, b₃] is transformed into [b₁, b₂, b₃, b₀].
- SubWord achieves a byte substitution on each byte of its input word, using Sbox.
- The result of steps 1 and 2 is XORed with a Round constant (Rcon[j])

The round constant is a word in which the three rightmost bytes are always 0. The round constant is different for each round and is defined as $Rcon[j] = (RC[j], 0, 0, 0)$, with $RC[1] = 1$, $RC[j] = 2 \cdot RC[j-1]$ and with multiplication defined over the field $GF(2^8)$. The values of $RC[j]$ in hexadecimal are listed in Table 2

Table 2 Round Constant Values

j	1	2	3	4	5	6	7	8	9	10
RC[j]	01	02	04	08	10	20	40	80	1B	36

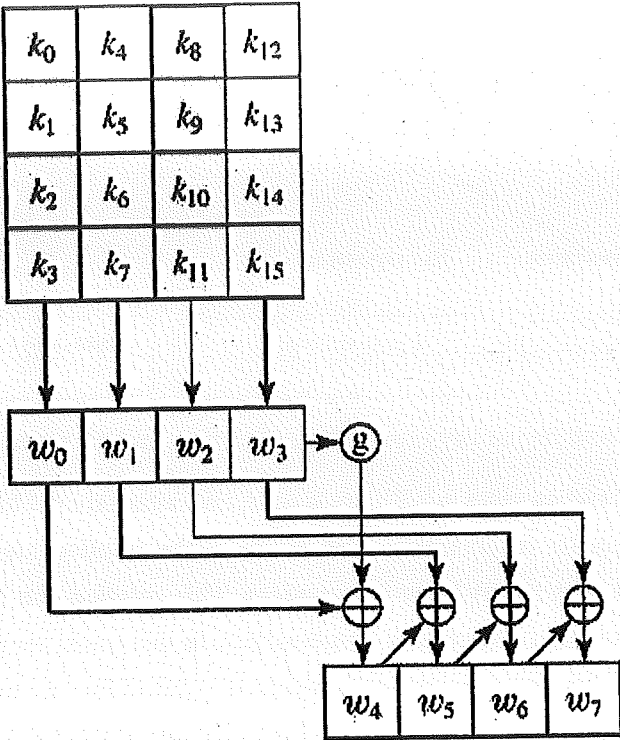


Figure 7 AES Key Expansion

III. Dynamic Rotation AES (DRAES)

The main purpose of rotation is to mix all data elements in different columns of state. As such, rotation is important for confusion and diffusion [8], which both plays an essential role in cryptography. Confusion refers to making the output dependent on the key. Ideally, every key bit influences every output bit.

Diffusion is making the output dependent on previous input (plain and ciphertext). Ideally, every previous input bit influences each output bit. One aim of confusion is to make it very hard to find the key even if one has a large number of plaintext-ciphertext pairs produced with the same key. Therefore, each bit of the ciphertext should depend on the entire key and in different ways on different bits of the key.

Rotation in AES is used in encryption and decryption as summarized in Figure 8. The encryption includes rotation in key expansion (cyclic shift-left of the round key bytes) and AddRoundKey (shift-left of state rows). The decryption includes inverse of the AddRoundKey (shift-right of state rows).

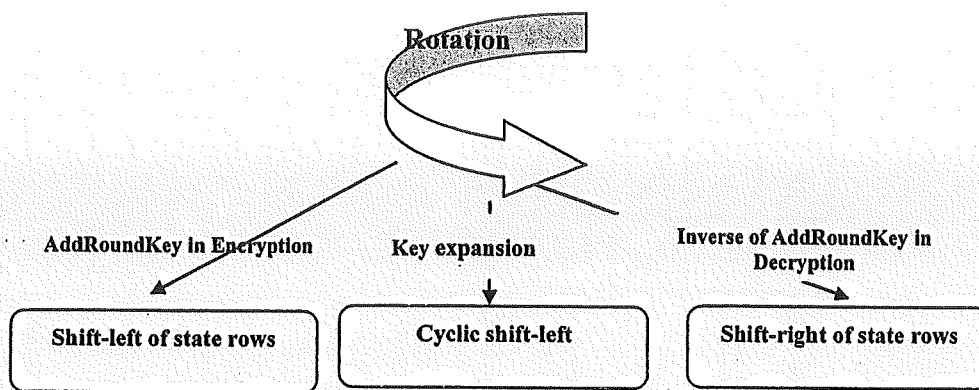


Figure 8 Rotation in key expansion, encryption, and decryption in AES.

The standard rotation rules in AES are simple, and attackers get high probability to break AES system by cryptanalysis. For example, the rotation amount (number of shifts) in AES is constant and key-independent. This rotation can be predicted and makes AES potentially vulnerable to cryptanalysis. In this work, we propose Dynamic Rotation for Advanced encryption Standard (DRAES), a modification and enhancement of the rotation in AES as follows. The proposed rotation enhancement makes the rotation amount dependent on data (plaintext and ciphertext) in AddRoundKey and on the key in key expansion. When rotation occurs the round key and intermediate data are modified, and either the key value or the intermediate data affect the rotation amount.

A) Key expansion rotation in DRAES

The following algorithm describes the proposed modification in the rotation inside the key expansion.

Constants: int Nb = 4;

Inputs: int Nk = 4, 6, or 8; // the number of words in the key

array key of $4 \cdot N_k$ bytes or N_k words // input key

Output: array W of $N_b \cdot (N_r + 1)$ words or $4 \cdot N_b \cdot (N_r + 1)$ bytes // expanded key

Algorithm:

1. $\text{int } N_r = N_k + 6;$
2. $W = \text{new byte}[4 \cdot N_b \cdot (N_r + 1)];$
3. $\text{int temp; int } i = 0;$
4. **while** ($i < N_k$) {
5. $W[i] = \text{word}(\text{key}[4 \cdot i], \text{key}[4 \cdot i + 1], \text{key}[4 \cdot i + 2], \text{key}[4 \cdot i + 3]);$
6. $i++;$ }
7. $i = N_k$
8. **While** ($i < N_b \cdot (N_r + 1)$)
9. **if** ($i \bmod N_k == 0$)
10. $W[i-1] = (b_{0,i-1}b_{1,i-1}b_{2,i-1}b_{3,i-1})$
11. $\text{temp} = (b_{0,i-1} \oplus b_{1,i-1} \oplus b_{2,i-1} \oplus b_{3,i-1})$
12. **if** ($\text{temp} \bmod N_k == 0$)
13. $W[i-1] = (b_{1,i-1}b_{2,i-1}b_{3,i-1}b_{0,i-1})$ % shift left by one byte to enforce rotation
14. **else if** ($\text{temp} \bmod N_k == 1$)
15. $W[i-1] = (b_{1,i-1}b_{2,i-1}b_{3,i-1}b_{0,i-1})$ % shift left by one byte
16. **else if** ($\text{temp} \bmod N_k == 2$)
17. $W[i-1] = (b_{2,i-1}b_{3,i-1}b_{0,i-1}b_{1,i-1})$ % shift left by two bytes
18. **else** ($\text{temp} \bmod 4 = 3$)
19. $W[i-1] = (b_{3,i-1}b_{0,i-1}b_{1,i-1}b_{2,i-1})$ % shift left by three bytes
20. **End if**
21. $\text{temp} = \text{SubWord}(W[i-1]) \wedge \text{Rcon}[i/N_k];$
22. **else if** ($N_k > 6 \ \&\& \ (i \bmod N_k) == 4$)
23. $\text{temp} = \text{SubWord}(\text{temp});$
24. $W[i] = W[i - N_k] \wedge \text{temp}; \quad i++; \quad \} \}$
25. **End while**

The RotWord rotation in key expansion occurs 10 times in DRAES similar to AES for key length of 128 bits ($N_k = 4$). Table 3 and Figure 8 show a comparison between AES and DRAES.

Table 3 Comparison between key expansion in DRAES and AES.

	DRAES	AES
Number of rotations	10	10
Maximum rotation amount per round	3	1
Variable rotation amount	Yes	No
Confusion and diffusion	high	Low
predictable	weak	rise

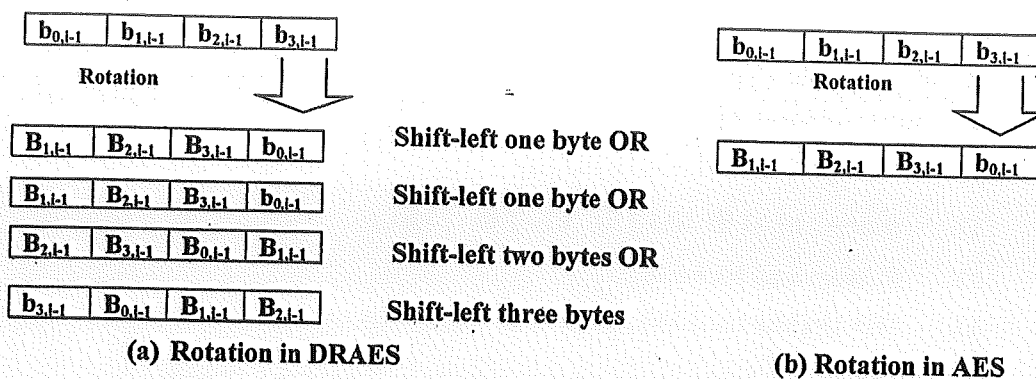


Figure 8 (a), (b). Rotation in key expansion for DRAES and AES

B) AddRoundKey rotation in DRAES

The modification of rotation in the ciphering process is vital; the change from constant shift-row to variable shift-row make the rotation amount hard to guess, which increases confusion and diffusion. In AES, row 0 is not shifted, row 1 is shifted 1 byte, row 2 is shifted 2 bytes, and row 3 is shifted 3 bytes. In DRAES, rotation amount is variable and done with the following procedure.

1. $Nb=4$ // the number of columns is denoted by Nb and is equal to the block length divided by 32
2. State $[4, Nb]$ // the State can be pictured as array of bytes. This array has four Rows, the number of columns is denoted by Nb and is equal to the block length divided by 32
3. $NR=10$ //number of round for key 128 bit, $NR=12$ for key 192 bit and $NR=14$ for key 256 bit
4. AddRoundKey(state, Roundkey)
5. For round = 1 step 1 to $Nr-1$
6. SubBytes (state, s_box)

7. ShiftRows (state) // the rotation for each row individually in state

I. read each row in state

II. Sum the elements in each row in Temp using Xor

III. If (Temp mod Nb=0)

IV. Shift left by one-byte to enforce rotation

V. else if (Temp mod Nb =1)

VI Shift left by one byte

VII. Else if (Temp mod Nb =2)

VIII. Shift left by Two byte

IX. Else (Temp mod Nb =3)

X. Shift left by three bytes

XI. // end if and end of ShiftRows (state)

8. MixColumns (state)

9. AddRoundKey (state, RoundKey)

10. // end for

11. SubBytes (state) // final round state

12. ShiftRows (state)

I. read each row in state

II. Sum the elements in each row in Temp using Xor.

III. If (Temp mod Nb =0)

IV. Shift left by one byte // to enforce rotation

V. else if (Temp mod Nb =1)

VI Shift left by one byte

VII. Else if (Temp mod Nb =2)

VIII. Shift left by Two byte

IX. Else (Temp mod Nb =3)

X. Shift left by three bytes

XI. // end if and end of ShiftRows (state)

13. AddRoundKey (State, RoundKey); 14. End cipher

This figure 9 and 10 illustrate the Rotation in AES and DRAES

C) DRAES in inverse cipher

The rotation in inverse cipher is the same process for the DRAES In cipher that described in sec. b Except for the shift row instead of shift row left, the shift row is right. Table 4 explain the variation between DRAES and AES for cipher

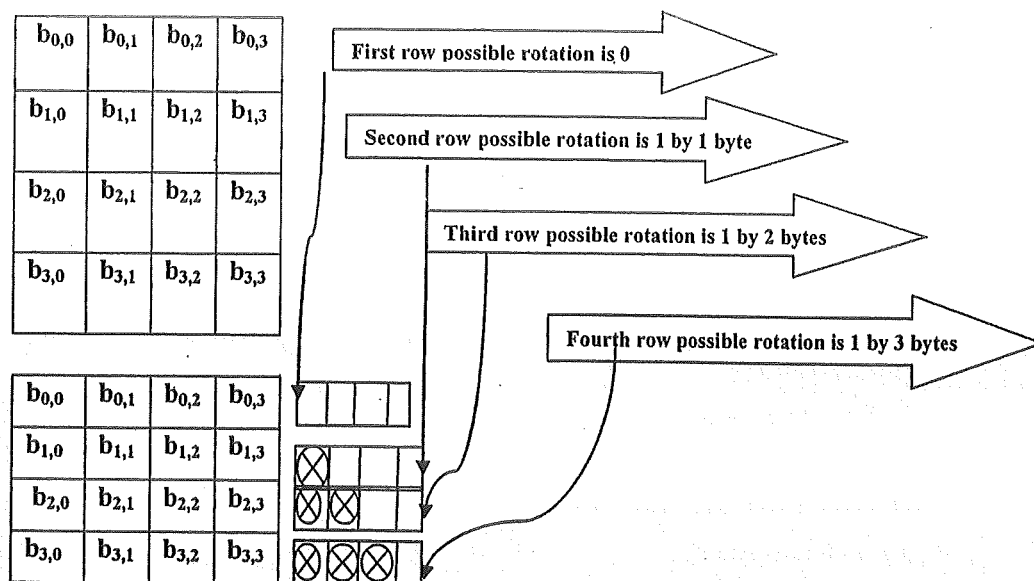


Figure 9 Possible Rotation (shift row left) states in (AES)

Table 4 demonstrate the conclusion between DRAES and AES for cipher

	DRAES	AES
number of row rotation for state in one round	4	3
number of row rotation for state in all round	36	27
number of row rotation for state in cipher	40	30
possible rotation for row 0 in state	$b_{0,1}$ $b_{0,2}$ $b_{0,3}$ $b_{0,0}$ $b_{0,1}$ $b_{0,2}$ $b_{0,3}$ $b_{0,0}$ $b_{0,2}$ $b_{0,3}$ $b_{0,0}$ $b_{0,1}$ $b_{0,3}$ $b_{0,0}$ $b_{0,1}$ $b_{0,2}$	$b_{0,0}$ $b_{0,1}$ $b_{0,2}$ $b_{0,3}$ $b_{0,0}$ $b_{0,1}$ $b_{0,2}$ $b_{0,3}$ $b_{0,0}$ $b_{0,1}$ $b_{0,2}$ $b_{0,3}$ $b_{0,0}$ $b_{0,1}$ $b_{0,2}$ $b_{0,3}$
possible rotation for row 1 in state	$b_{1,1}$ $b_{1,2}$ $b_{1,3}$ $b_{1,0}$ $b_{1,1}$ $b_{1,2}$ $b_{1,3}$ $b_{1,0}$ $b_{1,2}$ $b_{1,3}$ $b_{1,0}$ $b_{1,1}$ $b_{1,3}$ $b_{1,0}$ $b_{1,1}$ $b_{1,2}$	$b_{1,1}$ $b_{1,2}$ $b_{1,3}$ $b_{1,0}$ $b_{1,1}$ $b_{1,2}$ $b_{1,3}$ $b_{1,0}$ $b_{1,1}$ $b_{1,2}$ $b_{1,3}$ $b_{1,0}$ $b_{1,1}$ $b_{1,2}$ $b_{1,3}$ $b_{1,0}$
possible rotation for row 2 in state	$b_{2,1}$ $b_{2,2}$ $b_{2,3}$ $b_{2,0}$ $b_{2,1}$ $b_{2,2}$ $b_{2,3}$ $b_{2,0}$ $b_{2,2}$ $b_{2,3}$ $b_{2,0}$ $b_{2,1}$ $b_{2,3}$ $b_{2,0}$ $b_{2,1}$ $b_{2,2}$	$b_{2,1}$ $b_{2,2}$ $b_{2,3}$ $b_{2,0}$ $b_{2,1}$ $b_{2,2}$ $b_{2,3}$ $b_{2,0}$ $b_{2,1}$ $b_{2,2}$ $b_{2,3}$ $b_{2,0}$ $b_{2,1}$ $b_{2,2}$ $b_{2,3}$ $b_{2,0}$

possible rotation for row 3 in state	$b_{3,1}$	$b_{3,2}$	$b_{3,3}$	$b_{3,0}$	$b_{3,1}$	$b_{3,2}$	$b_{3,3}$	$b_{3,0}$
	$b_{3,1}$	$b_{3,2}$	$b_{3,3}$	$b_{3,0}$				
	$b_{3,2}$	$b_{3,3}$	$b_{3,0}$	$b_{3,1}$				
	$b_{3,3}$	$b_{3,0}$	$b_{3,1}$	$b_{3,2}$				
confusion and diffusion	high				Low			
predictable	hard				Easy			

VI. DRAES with confusion and diffusion

A strong cipher should contain both Confusion and diffusion. Claude Shannon, develop this concepts [9]. Confusion and diffusion are two techniques that symmetric ciphers should satisfy to thwart cryptanalysis. In a block cipher with good diffusion, if one bit of the plaintext digit is changed, then affects many cipher text digits in a random mode. Cryptographic diffusion test is a kind of statistical test that evaluates a block cipher for diffusion. The performance analysis can be done with various measures such as Diffusion analysis of DRAES and AES

VII. Diffusion Analysis

Diffusion makes the ciphertext dependent on previous plaintext and ciphertext. Diffusion is important for any block cipher, more specifically AES and DRAES algorithms. The impact of diffusion can be measured by the Strict Avalanche Criterion (SAC) [10], which is satisfied when at least 50% of bits in the ciphertext are changed in response to a one-bit flip in the plaintext or key.

Table 5 shows the SAC for both DRAES and AES when changing a single bit of plaintext while keeping the key constant. Table 6 shows the SAC for both DRAES and AES when changing a single bit of key while keeping the plaintext constant. Table 7 shows the SAC for both DRAES and AES when changing 3 bits of plaintext while keeping the key constant. Table 8 shows the SAC for both DRAES and AES when changing 3 bits of key while keeping the plaintext constant.

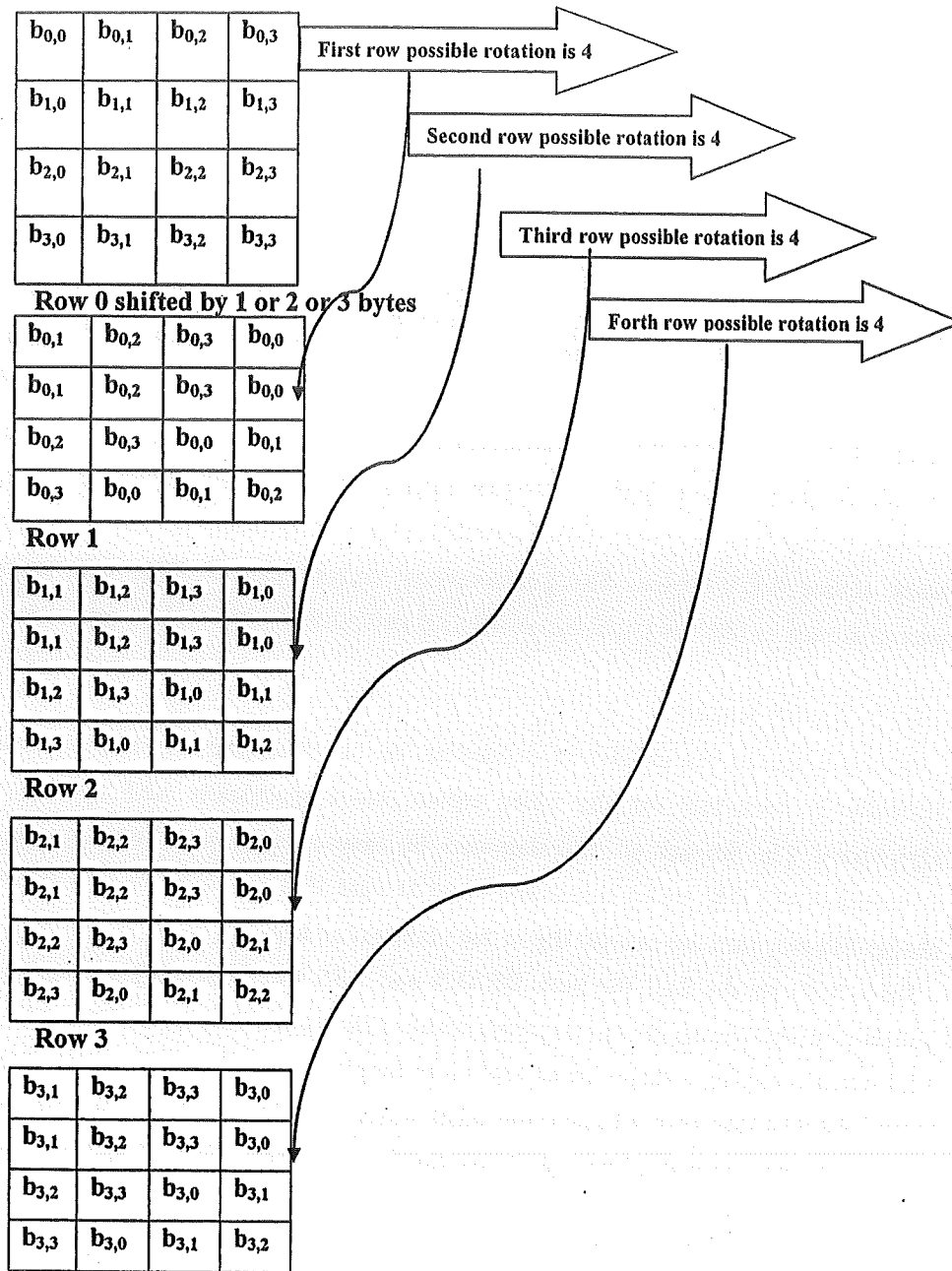


Figure 10 Possible rotation (shift row left) states in (DRAES)

Table 5 (a) (b) Avalanche effect for DRAES and AES
(Change in Plaintext by 1 bit & Key constant)

Round	Number of bit altered		SAC
1	11	9%	N
2	50	40%	N
3	77	61%	Y
4	59	47%	N
5	60	47%	N
6	69	54%	Y
7	63	50%	Y
8	65	51%	Y
9	60	47%	N
10	66	52%	Y

(a) AES

Round	Number of bit altered		SAC
1	16	13%	N
2	72	57%	Y
3	61	48%	N
4	63	50%	Y
5	63	50%	Y
6	64	50%	Y
7	83	65%	Y
8	61	48%	N
9	63	50%	Y
10	60	47%	N

(b) DRAES

As shown in Table 5 at first round, 11 bits (AES) of cipher value have changed out of 128-bit cipher text. This results an Avalanche value of 9%. SAC is achieved at the end of third round and Avalanche values transforms around the SAC value for the remaining rounds. Similar the Avalanche effect for (DRAES), the first round 16 bit of 128-bit with Avalanche value is 13%. SAC is achieved at the end of second round and Avalanche values changes around the SAC value more rapidly in DRAES than AES

Table 6 (a) (b) the Avalanche effect for DRAES and AES
(Alter 1 bit of key & the plaintext constant)

Round	Number of bit altered		SAC
1	36	29%	N
2	63	50%	Y
3	66	52%	Y
4	55	43%	N
5	72	57%	N
6	65	51%	Y
7	54	43%	N
8	63	50%	Y
9	63	50%	Y
10	66	52%	Y

(a) AES

Round	Number of bit altered		SAC
1	64	50%	Y
2	60	47%	N
3	69	54%	Y
4	61	48%	N
5	67	53%	Y
6	58	46%	N
7	68	54%	Y
8	66	52%	Y
9	70	55%	Y
10	67	53%	Y

(b) DRAES

The conclusion result in Table 6 demonstrate Avalanche effect SAC is achieved more rapidly in DRAES than AES in first round with SAC 50%, while the SAC for AES is completed in second round

**Table 7 (a) (b) details the Avalanche effect for DRAES and AES
(Varying 3 bits of plaintext & key constant)**

Round	Number of bit altered		SAC
1	28	22%	N
2	65	51%	Y
3	62	49%	N
4	57	45%	N
5	61	48%	N
6	55	43%	N
7	58	46%	N
8	63	50%	Y
9	67	53%	Y
10	55	43%	N

(a) AES

Round	Number of bit altered		SAC
1	28	22%	N
2	70	55%	Y
3	57	45%	N
4	73	58%	Y
5	64	50%	Y
6	68	54%	Y
7	67	53%	Y
8	61	48%	N
9	63	50%	Y
10	68	54%	Y

(b) DRAES

The end result in Table 7 display Avalanche effect SAC is achieved for DRAES and AES in same second round, but SAC 55% for DRAES is greater than SAC for AES.

**Table 8 (a) (b) illuminate the Avalanche effect for DRAES and AES
(Change in Key by 3 bits & Plaintext constant)**

Round	Number of bit altered		SAC
1	58	58%	Y
2	57	57%	Y
3	66	66%	Y
4	61	61%	Y
5	56	56%	Y
6	61	61%	Y
7	55	55%	Y
8	63	63%	Y
9	65	65%	Y
10	57	57%	Y

(a) AES

Round	Number of bit altered		SAC
1	76	60%	Y
2	65	51%	Y
3	63	50%	Y
4	66	52%	Y
5	73	58%	Y
6	71	56%	Y
7	66	52%	Y
8	60	47%	N
9	63	50%	Y
10	63	50%	Y

(b) DRAES

The outcome in Table 8 present Avalanche effect SAC is achieved for DRAES and AES in same first round, but SAC 60% for DRAES is greater than SAC for AES and greater in number of bits are altered (666 bits) than AES (599 bits).

V. CONCLUSION

With Dynamic rotation for advanced encryption standard DRAES the confusion and diffusion is stronger than rotation that occur in AES, that mean that Rijndael is more safe and physically powerful with dynamic rotation when compared to Rijndael with constant rotation as shown from results from tables that related with diffusion analysis

REFERENCES

- [1] Bernstein, Daniel J. "Understanding brute force." In *Workshop Record of ECRYPT STVL Workshop on Symmetric Key Encryption, eSTREAM report*, vol. 36, 2005.
- [2] Kumar, Neeraj. "Investigations in Brute Force Attack on Cellular Security Based on Des and Aes." *IJCEM International Journal of Computational Engineering & Management* 14 (2011).
- [3] Biryukov, Alex, Dmitry Khovratovich, and Ivica Nikolic. "Distinguisher and Related-Key Attack on the Full AES-256." *Advances in Cryptology-CRYPTO 2009: 29th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 16-20, 2009, Proceedings*. Vol. 5677. Springer, 2009.
- [4] Barkan, E., Biham, E.: In How Many Ways Can You Write Rijndael. In Zheng, Y., ed.: *Advances in Cryptology – ASIACRYPT 2002: 8th International Conference on Theory and Application of Cryptology and Information Security*. Volume 2501 of LNCS., Springer-Verlag (2002) 160–175
- [5] Krishnamurthy, G. N., and V. Ramaswamy. "Making AES stronger: AES with key dependent S-box." *IJCSNS International Journal of Computer Science and Network Security* 8, no. 9 (2008): 388-398.
- [6] Lu, Jiqiang, Orr Dunkelman, Nathan Keller, and Jongsung Kim. "New impossible differential attacks on AES." In *Progress in Cryptology-INDOCRYPT 2008*, pp. 279-293. Springer Berlin Heidelberg, 2008.
- [7] Serge Vaudenay, "A CLASSICAL INTRODUCTION TO MODERN CRYPTOGRAPHY", Springer Science+Business Media, Inc.,2006, ISBN-13: 978-0-387-25464-7
- [8] G. Lokeshwari, Dr. S. Udaya Kumar and G. Aparna "A CONFIGURABLE SECURED IMAGE ENCRYPTION TECHNIQUE USING 3D ARRAY BLOCK ROTATION", *International Journal of Engineering Science and Technology (IJEST)*, Vol. 4 No.01 January 2012
- [9] Drakakis, Konstantinos, Verónica Requena, and Gary McGuire. "On the Nonlinearity of Exponential Welch Costas Functions." *IEEE TRANSACTIONS ON INFORMATION THEORY* 56, no. 3 (2010).
- [10] Mohan H. S. and A Raji Reddy , "Performance Analysis of AES and MARS Encryption Algorithms", *IJCSI International Journal of Computer Science Issues*, Vol. 8, Issue 4, No 1, July 2011

1. The first part of the paper is devoted to a general discussion of the problem of the existence of a solution of the system of equations

which is the system of equations of the theory of the motion of a particle in a magnetic field.

2. In the second part of the paper the problem of the existence of a solution of the system of equations

is considered. It is shown that the system of equations has a solution if and only if the conditions

are satisfied. The conditions are satisfied if and only if the conditions

are satisfied. The conditions are satisfied if and only if the conditions

are satisfied. The conditions are satisfied if and only if the conditions

1871

1872

1873

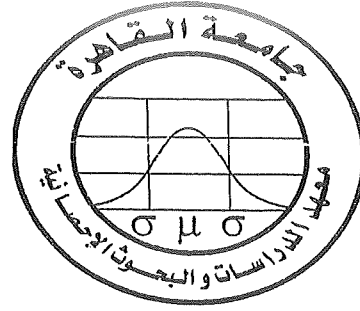
1874

1875

1876

1877

1878



جامعة القاهرة
معهد الدراسات والبحوث الإحصائية

المؤتمر السنوى السابع والأربعون

للإحصاء

وعلوم الحاسب

وبحوث العمليات

علوم حاسب

٢٤-٢٧ ديسمبر ٢٠١٢