Arabic Roots Extraction Using Morphological Analysis

Aymen Abu-Errub[1], Ashraf Odeh[2], Qusai Shambour[3], Osama Al-Haj Hassan[4]

**[1] Networks and Information Security Department, Al-Ahliyya Amman University,
Amman, Jordan**

**[2] Computer Information Systems Department, Isra University,
Amman, Jordan**

**[3] Software Engineering Department, Al-Ahliyya Amman University,
Amman, Jordan**

**[4] Computer Networks Department, Isra University,
Amman, Jordan**

Abstract

The Arabic language is characterized by its rich and complex morphology based on root-pattern schemes. Root extraction is one of the most important topics in the context of natural language processing applications such as information retrieval, text processing, machine translation, speech tagging, etc. This paper presents a method to extract the trilateral roots of Arabic words, acting from the roots of three consonants, through the removal of the prefixes and the suffixes, and the use of a list of morphological weights. Experimental results based on a list of eleven different root inflections shows the effectiveness of the proposed method with a success rate of 94%.

Keywords: *Arabic language, root extraction, stemming, morphological analysis.*

1. Introduction

The Arabic language is one of the oldest known spoken languages as well as one of the official languages of the United Nations. It belongs to the Semitic language family originated in the Arabian Peninsula in pre-Islamic times, and spread rapidly across the Middle East (Alatabbi & Iliopoulos,2012). According to the Internet World Stats (http://www.internetworldstats.com/stats7.htm), Arabic is the fifth most used language in the world, spoken by almost 340 million people in 27 states. The number of Arab Internet users in May, 2011 was about 65 millions which represents about 19 % of the population of the Arab world. Also, the Internet World Stats has reported that the number of Arabic speaking Internet users has grown 2,501.2 % in the last decade (2000-2011), which is the highest growth rate among other languages.

The Arabic language is very interesting in terms of its history, the strategic value of Arabic spoken people and the region they live in, and its cultural legacy. Historically, for more than fifteen centuries, classical Arabic remained unaffected, comprehensible and functional. At the Strategic level, Arabic is the native language of almost 340 million speakers occupying a main region with vast oil reserves important to the world economy. Culturally, the Arabic language is closely associated with Islam in which 1.4 billion Muslims perform their prayers five times daily (Farghaly & Shaalan,2009). In addition, the Arabic language is considered as a very challenging language due to: its complex linguistic structure in which it is characterized by a complex Diglossia situation, and its highly derivational nature where morphology plays a very important role (Farghaly & Shaalan,2009).

Morphology in linguistics is the study of the internal structure and formation processes of words. A morpheme is defined as the smallest meaningful and significant unit of language, which cannot be further broken down into smaller parts. We presents an example to clarify the complex nature of morphology in Arabic language where the style of writing letters in a word varies depending on the position of the letter within the word. Thus, the letter shapes changes if the letters comes at the beginning, middle or at the end of the word. In English, most of the changes take place at the beginning and the end of the word, leaving the core untouched. Arabic, in other hand, adds letters, or combinations of letters, between the root letters, as well as on the beginning and end. Accordingly, many definite articles, conjunctions, particles and other prefixes can be appended to the beginning of a word, and large numbers of suffixes can be attached to the end of a word. In addition, most noun, adjective, and verb stems are derived from a few thousand roots by infixing. For example, such words like (مكتب maktab) which means (a desk/an office), (كتاب kitaab) which means (a book),(كُتُب kutub) which means (books), (نكتُب naktubu) which means (we write), and (مكتبة maktaba) which means (a library), come from the root (كتب ktb). Stemming is the process of reducing the derived words to their base form (Abu Hawas,2013; Jurafsky & Martin,2009; Wightwick & Gaatar,2008).

The Arabic morphological analysis has been extensively used in several domains of the Arabic natural language processing (ANLP) such as Information Retrieval (IR) systems, electronic dictionaries, text classification, text summarization, text mining, etc (Farghaly & Shaalan,2009; Iazzi et al.,2013; Yousfi,2010). Darwish (2002) classifies the root extraction (stemming) strategies into three main approaches: the symbolic approach, the statistical approach, and the hybrid approach. The symbolic approach (also referred to as rule-based approach) is based on the linguistic rules that governing the combination of morphemes to segment the Arabic word into prefixes, infixes and suffixes in order to extract the root. The statistical approach calculates the probabilities that a prefix and a suffix or a radical can come out together in a database of words. The hybrid approach integrates the two previous approaches (Darwish,2002; Iazzi et al.,2013; Yousfi,2010).

Many algorithms have been proposed and implemented to extract Arabic roots, however, with limited success (Hmeidi et al.,2010). Accordingly, much more research is needed to further develop and refine the area of Arabic roots extraction. In the study, the researchers will introduce an improved symbolic approach to extract the word's root with the use of morphological rules and weights, which simplifies the process of extracting the roots. The rest of this paper is organized as follows: In Section 2, a brief overview of the related studies in which a number of research papers that deal with extracting Arabic word's roots are presented. Section 3 demonstrates the proposed algorithm. Section 4 presents the experimental setup and discuses the results. Finally, conclusions and directions for future study are provided in Section 5.

2. Previous Studies

Among the successful approaches for extract Arabic words roots, a number of recent studies on Arabic morphological analysis have been proposed (Abu Hawas,2013; Alatabbi & Iliopoulos,2012; Altantawy et al.,2010; Dilekh & Behloul,2012; Iazzi et al.,2013; Khorsi,2012; Yousfi,2010).

In the article presented by Yousfi (2010), a new approach of morphological analysis using the surface patterns of the word to be analysed is proposed. It is able to analyze all Arabic verbs by decomposing the word to the prefixes, suffixes and roots. The proposed approach remains always true for the derivative names and it is adequate to construct a database of surface patterns of derivative names. Finally, the approach has been evaluated on a corpus of 4000 verbs, the error rate is 4% and the duration of these verbs is 10 seconds (2.5 ms for each verb).

Altantawy et al. (2010) have extended their previous version of the MAGEAD system (Habash & Rambow,2006), which is a morphological analyzer and generator for Modern Standard Arabic (MSA) and Levantine Arabic verbs, to model MSA nominals (nouns and adjectives) that are far more complex to model than verbs. In their recent paper (Altantawy et al.,2010), they present the details of an implementation of MSA nominal in MAGEAD. A detailed evaluation of the current implementation comparing it with a commonly used morphological analyzer shows that it has good coverage and usability with high precision and recall. An error analysis reveals that the majority of recall and precision errors are problems in the gold standard or a result of the inconsistency between different models of form-based/functional morphology.

Khorsi (2012) presented a completely unsupervised approach for the extraction of morphological segments from classical Arabic words. The stemming is a preparatory step to an unsupervised root (i.e., radicals) extraction. As a learning input, the proposed stemming system requires no linguistic knowledge but a plain classical Arabic text, and is able to extract the strongest segment of a given length, specifically the stem, once the learning input is analyzed. The system performs about 90% true positives after a leaning of less than 15000 words. The proposed system, unlike other unsupervised approaches, does not presume the idealness of the input text and deals efficiently with the eventual (practically very frequent) misspellings. The test corpus used is an ultimate reference in the classical Arabic and its labelling has been thoroughly done by a group of experts.

Dilekh and Behloul (2012) proposed a hybrid method of stemming by integrating three different techniques (deletion of affix, dictionaries, and morphological analysis) to improve the overall performance of the stemming process. These techniques are applied individually and independently to solve associated stemming problems. The authors develop an information retrieval system dedicated to Arabic language based on the proposed hybrid method in the phase of stemming. Empirical results demonstrate the effectiveness of the hybrid method compared to other methods, and the choice of removing the suffix before prefix during the operation of Arabic stemming process.

Alatabbi and Iliopoulos (2012) developed a software, called as ‘AMA’, based on a new stemming algorithm. The proposed stemming algorithm is simple and highly effective in which it can reduce stemming errors, reduce computational time and data storage. The proposed algorithm uses the grammatical rules of Arabic language to achieve the goal of finding the root of a word in two steps, first it finds the word lemma then, based on the lemma, it determines the root of the word.

In Abu Hawas (2013), the author developed a new root-extraction approach for Arabic words that attempts to assign a unique root for each Arabic word without having an Arabic roots list, a words patterns list, or the Arabic word’s prefixes and suffixes list. The proposed approach predict the letters positions that may form the word root one by one, using rules based on the relations between the Arabic word letters and their placement in the word. The proposed approach consist of two parts, the first part deals with the rules that distinguish between the Arabic definite letter “ال AL, La” and the original word letters “الـ“. The second part of the approach adopts the segmentation of the word into three parts and classifies its letters into groups according to their positions. Empirical experiments of the proposed approach using the Holy Quran words show significant results in handling the word root.

Iazzi et al.(2013) introduced a morphological approach for the analysis of the Arabic language. The proposed approach is based on the surface patterns of Arabic words to deal with Arabic derived nouns. It is based mainly on the construction of a database for the surface patterns of the Arabic derived nouns. Experiments were conducted against a corpus of 2400 Arabic words (400 verbs and 2000 derived nouns), and the error rate found was reasonable. In general, the obtained results demonstrate the effectiveness of the proposed approach.

3. Proposed Method

This section presents the proposed method to extract the roots of words. First, this method uses stemming algorithm to remove suffixes and prefixes that are added to the root, then the infixes are deleted through comparisons with the morphological inflections. The proposed method is characterized by not using any special roots to compare with, but based on a series of steps to find the root. Following is an explanation of the proposed method:

3.1 Stemming

The first step of the proposed method is "stemming" process, which aims to delete the word's prefixes and suffixes letters.

3.1.1 Deleting Prefixes Letters

The process is started by deleting the prefixes letters of the word. After the deletion, the number of word's letters is counted to make sure that there are still more than three letters. If the number of letters after the deletion is less than three, the deletion process is canceled (as shown in Fig. 1). Following are the detailed steps:

1. Define the following variables:

Input word	W
Input word length	N
Prefixes matrix	P
Prefix length	Li
Temporary value	temp
Result	R
Length of word	length

2. Create a list of prefixes matrix (P), as shown in Table 1.
3. Determine the length of input word and store the result in (N).
4. If (N) is less than or equal to 3 then mark (R = W) go to step 10.
5. Start deleting the prefixes by reading the prefix from the prefixes matrix (P).
6. If any prefix elements matches with the beginning of the word (W) delete the prefix (Li) from the beginning of the word and store the output in (temp).
7. If the length of (temp) is less than 3, cancel the deletion process and move to the next prefix in the prefixes matrix, and repeat step 4.
8. If the length of (temp) is more than 3, let R = temp and go to Step 5.
9. If the prefix matrix (P) element does not match with the beginning of the word (W), move to the next prefix and repeat step 4 until end of prefix matrix (P).
10. Return R.

Table 1: Prefixes Matrix (P)

فال	بال	فلل	سيس	مست
فكال	فبال	كال	سيست	ستس
تست	ال	با	كا	سا
اف	است	ستت	سيت	يست
فب	فك	ست	فل	اس
ف	يتس	و	ي	لل

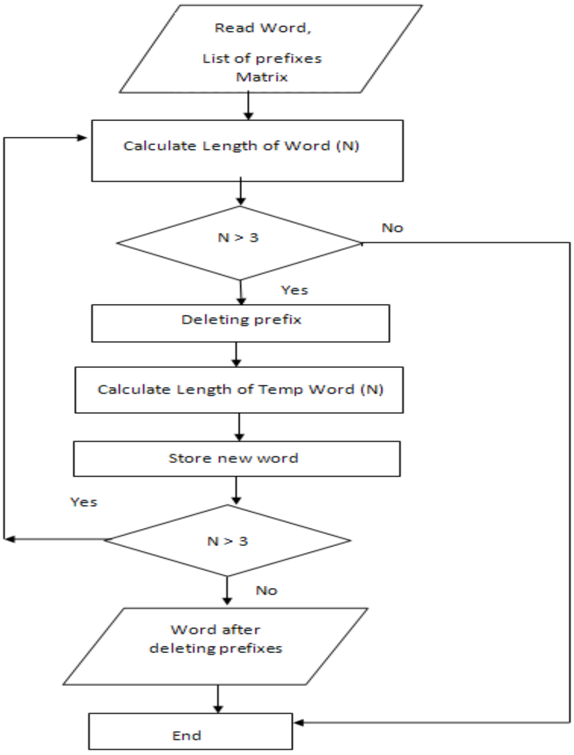
Read Word,
List of prefixes Matrix
Calculate Length of Word (N)
N > 3
No
Yes
Deleting prefix
Calculate Length of Temp Word (N)
Store new word
Yes
N > 3
No
Word after deleting prefixes
End

Fig. 1 Deleting Prefix Letters

3.1.2 Deleting Suffixes Letters

The process of deleting the prefix letters is repeated here for deleting the suffix letters. And after the deletion is completed, the number of the word's letters is counted to make sure that they are still more than three letters. If the number of letters after the deletion is less than three, we cancel the deletion process. Following are the deletion process steps (as shown in Fig. 2):

1. Define the following variables:

Input word	W
Input word length	N
Suffixes matrix	S
Suffix length	Mi
Temporary value	temp
Result	R
Length of word	length

2. Create a list of suffixes, shown in Table 2.
3. Determine the length of input word and store the result in (N).
4. If (N) is less than or equal to 3 Then Mark (R = W) go to Step 10
5. Start deleting the suffixes by reading the suffix from the suffixes matrix (S).
6. If any suffixes matrix (S) elements matches with the end of the word (W) Delete the suffix (Mi) from the end of the word and store the output in (temp).
7. If the length of (temp) is less than 3, cancel the deletion process and move to the next suffix in the suffixes matrix (S), and repeat step 4.
8. If the length of (temp) is more than 3, let R = temp and go to Step 5.
9. If the suffixes matrix (S) element does not match with the ending of the word (W), move to the next suffix and repeat step 4 until end of suffixes matrix (S).
10. Return R.

Table 2: Suffixes Matrix (S)

نا	كن	كم	ها	هن	هم
تم	يه	ه	كما	هما	تما
ون	ين	ان	وا	ات	تن

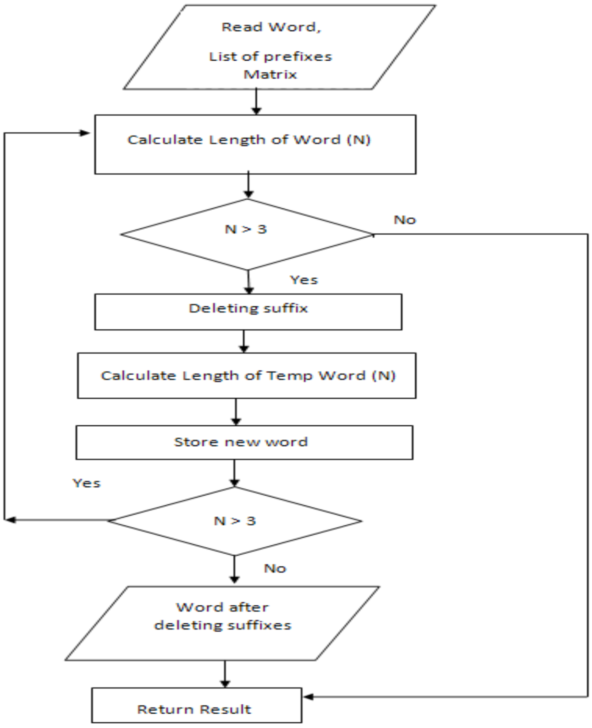

Fig. 2 Deleting Suffix Letters

After completing the stemming process, the resulted word will be without prefixes and suffixes, if the resulted word length is three letters, it is considered as the root, and otherwise it contains extra letters in the middle of the word, at the beginning or at the end.

The Arabic morphological weight consists of three letters (ف ع ل) , the first letter (ف) corresponds to the first letter of the root, and the letter (ع) corresponds to the second one, and the letter (ل) corresponds to the third letter, for example, the root, "كتب" the letter "ك" corresponds to the letter "ف" in the morphological weight, and the letter "ت" corresponds to the letter "ع" and the letter "ب" corresponds to the letter "ل". When you form the morphological inflection for any root, one letter or more are added to the root. This addition will take place on the word according to the number of letters and the place of the addition. Table 3 shows the addition on the words and the corresponding word it in the morphological weight.

Table 3: Inflection of Roots

Root	***Inflection***	***Resulted Word***
كتب	فاعل	كاتب
ركب	مفعل	مركب
غفر	استفعل	استغفر

3.1.3 Root Extracting Algorithm

To extract the root, the resulted word of the stemming process is to be compared with its appropriate morphological weight, and then the extra letters are deleted from the word, which correspond to the extra letters in the morphological weight. The following algorithm describes how to extract the root after Stemming, (as shown Fig 3):

1. Define The used symbols:

Morphological weights matrix	Awzan
Morphological weight	Mwi
Weight length	L
Input word	W
Input word length	N

2. Read the length of the word (W) and store it in (N).
3. Read the length of the first element of the Morphological weights matrix and store it in (L).
4. If the length of morphological weight (L) is equal to the length of the input word (N), compare the extra letters in the morphological weight with the corresponding letters in the input word.
5. If the extra letters in the morphological weight match with the letters in the input word delete them and return the rest of the word as the root.
6. If the extra letters in the morphological weight does not match the corresponding ones in the input word, move to the next morphological weight and repeat step 3.
7. If the length of the weight (L) does not equal to the length of the Input word (N), move to the next weight and repeat step 3.

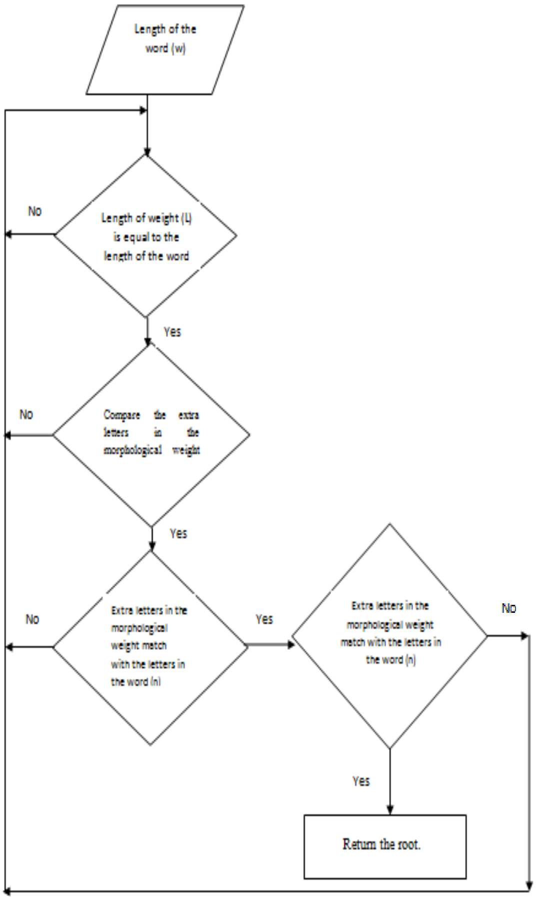
Length of the word (w)
No
Length of weight (L) is equal to the length of the word
Yes
No
Compare the extra letters in the morphological weight
Yes
No
Extra letters in the morphological weight match with the letters in the word (n)
Yes
Extra letters in the morphological weight match with the letters in the word (n)
No
Yes
Return the root.

Fig. 3 Root Extracting Algorithm

4. Experimental Results

To test the proposed algorithm we have created a special list (corpus) for examining purpose. The objective of this list is to make sure that we get the appropriate number of morphological inflections to test the proposed method. To form the corpus, one hundred roots are selected in a random manner, and 11 different inflections for each root are formed, the resultant list contains 1100 morphological inflections. Table 4 demonstrates the morphological inflections that are formed.

Table 4: Morphological Inflections

Passive participle	External plural
Internal plural	Instrument name
Pronouns' referring	Time and place Adverbs
Dual	Active participle
Superlative	Intensifying apposition with "non"
Feminine form	

The proposed algorithm was examined using the resultant corpus and the results were as shown in Table 5:

Table 5: Root Extraction Rate

Inflection	*Success rate*	*Correct roots*	*Wrong roots*
Active participle	%82	82	18
Passive participle	%93	93	7
Dual	%81	81	19
External plural	%85	85	15
Internal plural	%91	91	9
Instrument name	%96	96	4
Superlative	%79	79	21
Intensifying apposition	%75	75	25
Feminine form	%75	75	25
Pronouns' referring	%82	82	18
Time and Place Adverbs	%99	99	1
Total	%94	938	162

It's noted from the table that the best success rate of extracting the root was to the inflection of "Time and Place Adverbs" which equal to 99%, while the lowest percentage was to the both inflections of "Intensifying apposition" and "Feminine form", as it reached to 75%, while the overall percentage of all inflections was 94%. Fig. 4 illustrates the average of success and failure for some inflection**.**

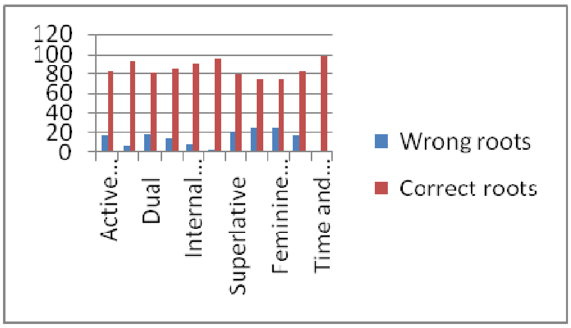

Fig. 4: The average of success and failure rate for some inflection

5. Conclusions

This paper presents a new algorithm for Arabic roots extraction using morphological analysis. The proposed algorithm uses stemming operation to prepare the word for use, and then it is compared to a list of different morphological weights. A list of different morphological weights to examine the proposed algorithm has been prepared to ensure selection of the maximum amount of morphological inflictions, after testing the contents of the proposed list the percentage of correct extraction roots using the proposed algorithm was 94%. The proposed algorithm needs more work and modification to be valid for extracting the defective roots, quadrilateral, fivefold, and six fold roots.

References

[1] Abu Hawas, F. (2013), "Exploit relations between the word letters and their placement in the word for arabic root extraction", *Computer Science*, vol. 14, no. 2, pp. 327-341.

[2] Alatabbi, A. & Iliopoulos, C.S. (2012), "Morphological analysis and generation for Arabic language", *International Conference on Computer Science, Engineering & Technology*, pp. 1-9.

[3] Altantawy, M., Habash, N., Rambow, O. & Saleh, I. (2010), "Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach", *Proceedings of the seventh International Conference on Language Resources and Evaluation*, Valletta, Malta, pp. 851-858.

[4] Darwish, K. (2002), "Building a shallow Arabic Morphological Analyzer in one day", *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, Association for Computational Linguistics, Philadelphia, Pennsylvania, pp. 1-8.

[5] Dilekh, T. & Behloul, A. (2012), "Implementation of a New Hybrid Method for Stemming of Arabic Text", *International Journal of Computer Applications*, vol. 46, no. 8, pp. 14-19.

[6] Farghaly, A. & Shaalan, K. (2009), "Arabic Natural Language Processing: Challenges and Solutions", *ACM Transactions on Asian Language Information Processing*, vol. 8, no. 4, pp. 1-22.

[7] Habash, N. & Rambow, O. (2006), "MAGEAD: a morphological analyzer and generator for the Arabic dialects", Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Sydney, Australia, pp. 681-688.

[8] Hmeidi, I.I., Al-Shalabi, R.F., Al-Taani, A.T., Najadat, H. & Al-Hazaimeh, S.A. (2010), "A novel approach to the extraction of roots from Arabic words using bigrams", *Journal of the American Society for Information Science and Technology*, vol. 61, no. 3, pp. 583-591.

[9] Iazzi, S., Yousfi, A., Bellafkih, M. & Aboutajdine, D. (2013), "Morphological Analyzer of Arabic Words Using the Surface Pattern", *International Journal of Computer Science Issues*, vol. 10, no. 2, pp. 254-258.

[10] Jurafsky, D. & Martin, J.H. (2009), Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd edition edn, Prentice Hall.

[11] Khorsi, A. (2012), "Effective Unsupervised Arabic Word Stemming: Towards an Unsupervised Radicals Extraction", *Int. Arab J. Inf. Technol.*, vol. 9, no. 6, pp. 571-577.

[12] Wightwick, J. & Gaatar, M. (2008), *Arabic Verbs and Essentials of Grammar*, McGraw-Hill, Chicago.

[13] Yousfi, A. (2010), "The morphological analysis of Arabic verbs by using the surface patterns", *International Journal of Computer Science Issues*, vol. 7, no. 3, pp. 33-36.

Aymen Abu-Errub is an Assistant Professor in Faculty of Information Technology in Al-Ahliyya Amman University, Jordan in Computer Information Systems department and then in Networks and Information Security departmentreceived. He received his Ph.D. degree in Computer Information Systems from the Arab Academy for Banking and Financial Sciences, Jordan, in 2009. He has published journal papers in information retrieval, information and networks security, and in risk management fields. He also participated and published his researches in scientific conferences.

Ashraf A. Odeh is an Assistant Professor in Computer Information System at Isra University-Jordan. He received a BSc degree in Computer Science in 1995 and MSc degree in Information Technology in 2003. With a Thesis titled "Visual Database Administration Techniques", He received PhD from department of Computer Information System in 2009 with a Thesis titled "Robust Watermarking of Relational Database Systems". He is interested in image processing, Watermarking, Relational Database, E-copyright protection, E-learning and E-content. He has submitted a number of conference papers and journals. Also he has participated in a number of conferences and IT days.

Qusai Shambour is an Assistance Professor in the department of Software Engineering, Faculty of Information Technology, at Al-Ahliyya Amman University, Jordan. He received his PhD from University of Technology Sydney (UTS), Australia in 2012. His main research interests lie in the area of information retrieval, web personalization, recommender systems and e-Government and e-Service intelligence. He has published 15 papers in refereed journals and conference proceedings.

Osama Al-Haj Hassan obtained his BS in Computer Science from Princess Sumayya University for Technology (PSUT). He also received his MS in Computer Science from New York Institute of Technology (NYIT). He obtained his PhD in Computer Science from University of Georgia (UGA). Currently, he is an Assistant Professor at Isra University in Jordan. His research interests are in distributed systems, web services, Web 2.0, caching and replication techniques, peer-to-peer networks and event based systems..