Arabic Stemmer for Search Engines Information Retrieval

Ahmed Khalid
Computer Science Department
Community College
Najran University
Najran, KSA

Zakir Hussain
Computer Science Department
Community College
Najran University
Najran, KSA

Mirza Anwarullah Baig
Computer Science Department
Community College
Najran University
Najran, KSA

***Abstract*—Arabic language is very different and difficult structure than other languages, that's because it is a very rich language with complex morphology. Many stemmers have been developed for Arabic language but still there are many weakness and problems. There is still lack of usage of Arabic stemming in search engines. This paper introduces a rooted word Arabic stemmer technique. The results of the introduced technique for six Arabic sentences are used in famous search engines Google Chrome, Internet Explore and Mozilla Firefox to check the effect of using Arabic stemming in these search engines in terms of the total number of searched pages and the search time ratio for actual sentences and their stemming results. The results show that Arabic words stemming increase and accelerate the search engines output.**

***Keywords*—Information Retrieval; Arabic Stemming; Search Engine; Arabic Morphology**

I. INTRODUCTION

Arabic language is the one of widely spoken language in the world [21]. It belongs to the Semitic languages branching family of the Asian-African languages Group [8]. The morphology of Arabic poses special challenges to computational natural language processing systems. The exceptional degree of ambiguity in the writing system, the rich morphology, and the highly complex word formation process of roots and patterns all contribute to making computational approaches to Arabic very challenging[6]. Arabic words are formed from abstract forms named roots, the root is the basic form of word from which many derivations can be obtained by attaching certain affixes or suffixes so we produce many nouns and verbs and adjectives from the same root [17,20].

Furthermore Arabic is a highly inflectional language with 85% of words derived from trilateral roots. Nouns and verbs are derived from a closed set of around 10,000 roots [27]. Arabic words are normally derived from bilateral, trilateral, quadri-literal, and pentaliteral verbs. These words represent various modifications of the original verbs. These modifications are represented by special types of measures called templates 'اوزان". Arab grammarians use the verb "فعل" 'faal" as the model to represent the templates that can form other verbs and nouns from a specific root. Each letter of the special verb "فعل" has a specific name and meaning applied in the formation of other forms of verbs and nouns, the first letter is called "ف" "fa", the second is "ع" "ain", and the third is "ل" "lam" [26]. Many stemmer techniques for Arabic language depend on root extraction [20,26]. One of the Arabic language characteristics the formats of the words in Arabic language are the union of templates for the meanings vary in function performed. "فالناظر" " Beholder ", "والمنظور" " Perspective " , "المنظر" "Theorist" vary in significance with the agreement at the root of the general concept, which is a "نظر "look".

Stemming has multiple definitions, it is the process of reducing a word to its stem or root form. Stemming is considered by number of authors as word standardization [25]. Shereen Khoja's define stemming as the process of removing all of a word's affixes to produce the stem or root. However, Leah Larkey [19] was more general in her definition stemming refers to any process which conflates related forms or groups forms into equivalence classes, including but not restricted to suffix stripping. Stemming might be useful to Information retrieval systems, text classification systems, text clustering systems, dictionary automation, text compression, etc.

Sharul classify Arabic stemming to four different approaches manually constructed dictionaries used by Al-Kharashi and Evens[4], algorithmic light stemmers which remove prefixes and suffixes which is mentioned by some authors [5,11], morphological analyses which attempt to find roots Several morphological analyzers have been developed for Arabic [3,6,7,10], and statistical stemmers, which group word variants using clustering techniques Statistical techniques have widely been applied to automatic morphological analysis in the field of computational linguistics [9,12,14,15,16].

There are many Stemming techniques for Arabic language , Al-Stem Stemmer developed by Kareem Darwish and modified by Leah Larkey from University of Massachusetts and further modified later by David Graff form LDC, in which (وال,فال ,بال,بت,يت,لت,مت,وت,ست,نت,يم,لم,وم,كم ,فم ,ال ,لل ,وي ,ل ,ي,سي ,في ,وا,فا,با) are rmoved from the prefixes and (ات,وا,ون,وه,ان,تي,ته,تم,كم,هن,هم,ها,يـ,تك,نا,ين,يه,ه,ي,ا) are removed from the suffixes of the word [22]. Aljlayl Stemmer [5] Mohammed Aljlayl developed a light stemmer used for his own information retrieval researches in TREC cross-language track in that stemmer he considered the length of the word for removing suffixes and affixes, also he normalize certain Arabic characters like أ , إ ,ا [23]. Light8 Stemmer[19] It is a light stemmer developed by Leah Larkey for the purpose of

researching, the stemmer remove و if the remainder of the word is 3 or more characters long. Remove any of the definite articles if this leaves 2 or more characters. Remove any of the following suffixes in order form right to left (ي,ها,ان,ات,ون,ين,يه,ه) If this leaves 2 or more characters. Light10 stemmer which is a modification of light8 stemmer. Berkeley light stemmer [2] is gives best performance in the track. This stemmer depends on the length of the word if the word is at least five-character long, remove the first three characters if they are one of the following (مال,اال,سال,لال,وال,بال,فال,كال,ولل). If the word is at least four-character long, remove the first two characters if they are one of the following (لل,فا,يا,سي,وم,وت,ال,وي,وا,لا,وب,ول,وس,كا). If the word is at least four-character long and begins with و, remove it. If the word is at least four-character long and begins with either ب or ل, remove ب or ل, only if, after removing the initial character, the resultant word is present in the Arabic document collection. Recursively strips the following two-character suffixes in the order of presentation if the word is at least four-character long before removing a suffix (ون,ات,ان,ين,تن,تم,كن,كم,هن,ني,يا,وا,ما,نا,هم,يو,ها). Recursively strips the following one character suffixes in the order of presentation if the character is at least three-character long before removing a suffix (ت, ي,ه). Kadri's linguistic-based stemmer[29] this stemmer depended on the idea that the Arabic word consists of five part their order is; antefixes, prefixes, stem, suffixes and postfixes. The linguistic-based stemmer has two phases, Training Phase a list of stems with its frequency occurrence is built for each word using corpus to avoid ambiguity that may happen when removing affixes. The Stemming Phase where the stemmer truncates possible affixes according to the above list. If there is an ambiguity raised for the stemmer (more than one combination was available), then stemmer selects the most appropriate candidate; according to corpus statistics computed in the training phase. Restrict Stemmer [24] focused on removing conjunctions, و, and ف, and prepositions, ك, و, ل , and ب, that come as prefixes in the beginning of the words. Here didn't mention other affixes such as articles and suffixes in general just tried to find a way to recognize the two types of affixes; the conjunctions and the prepositions. Beltagy Stemmer Samhaa El-Beltagy and Ahmed Rafea [13] have proposed a stemming technique that not only removes prefixes and suffixes from the beginning and the end of the word, but also converts the irregular plural form of the word to its singular form. The stemmer also is a domain specific stemmer which conducts stemming according to the domain of the collection of text to be indexed. The domain specific idea is implemented using a stem list that contains the words and their stems. So, before accepting a stem that produce from a word using stem-based stemmer, the system check whether the produced stem exist in the list or not.

This paper presents Arabic stemming techniques depends on the root of the Arabic word. six Arabic sentences and their stemming results from this technique are used to see the effect of the Arabic stemming usage in the famous search engine Google Chrome, Internet Explore and Mozilla Firefox in terms of the number of the web pages and the time taken by these search engine for the original sentences and the stemming results of this sentences.

The rest of this paper is organized as follows: section II presents the related works ; section III introduce the proposed Arabic stemming techniques, section IV shows the results and discussions; and finally section V concludes the work.

II. Related Works

Considerable research on stemming and morphological analysis is tedious for the Arabic language, but no standard IR-oriented algorithm has yet emerged. Furthermore there is a good effort done to achieve standard Arabic stemming algorithm, here we show the researchers efforts. Aitao Chen and Fredric Gey [2] developed one MT-based Arabic stemmer and one light Arabic stemmer. The Berkeley light stemmer worked better than the automatically created MT-based stemmer. The experimental results show query expansion substantially improved the retrieval performance. Mohamad Ababneh, Riyad Al-Shalabi, et [20] develop a rule-based light depends on truncation of affixes and introduced a new algorithm uses a set of rules to determine if a certain sequence of characters is part of the original word or not and this helped solving some ambiguity problems, also they introduced a way for handling the majority of broken plural forms and reducing them to their singular form which helped to group words of the same meaning in a common form. Kazem Taghva, Rania Elkhoury, Jeffrey Coombs [18] introduce an Arabic Stemming Without A Root Dictionary their experiment shows that stem lists are not required in an Arabic Stemmer. Riyad Al-Shalabi, et [26] build Arabic stemmer based on Excessive Letter Locations, The stemmer find the trilateral root, quadri-literal root as well as the pentaliteral root for Arabic words based on excessive letter locations. The algorithm locates. The goal of stemmer is to support natural language processing programs such as parsers and information retrieval systems.

III. Arabic Stemming Technique

The morphology complexity of Arabic makes it particularly difficult to develop natural language processing applications for Arabic information retrieval. Stemming is another one of many tools besides normalization that is used in information retrieval to combat this vocabulary mismatch problem. The proposed Arabic stemming algorithm for our research is constructed from the following process:

1) *Deleting of Arabic stop words*
2) *Striping of diacritics*
3) *Striping of affix (prefixes and suffixes)*
4) *Determining the stem*
5) *Recoding the stem*
6) *Verifying the stem with a dictionary of root words*

Step1: The proposed algorithm eliminates stop words from the document using the Arabic stop words list which contains 162 words like ..... , من, قد ' لم ' في

Step2: Normalize the rest of the words: such as removing punctuation, converting words to lowercase, stripping numbers out, as following:

- Remove punctuation.
- Delete numbers, spaces and single letters.

- Convert letters (إ) ,(أ) ,(ئ) ,(ؤ), (ء) into (ا), and (ة) into (ه).

Step3: Applying techniques to find the basic root of Arabic words by removing affixes (suffixes and prefixes) attached to its root. Prefix like (ال, وال, بال, كال، فال, لل) and suffix like (ي, ةو ه, ون, ين,ات, ية,يه, ان,ها)

Step4: The stemming process begins by processing a word and trying to find its correct stem. In case the word does have a correct stem, then the word without its affixes will be returned. The stemming algorithm will take as input an Arabic word (not a stop word), and the output will be the extracted root (or stem). In cases where the algorithm cannot find a root for the specific word, the word itself will be taken as a root. Such cases are few and it depends on the quality of the algorithm proposed

Step5: The recoding module main concern is to change some of the letters to their correct form. These changes will probably occur during the process of template formation when a word is formed from a root. Some letters may be dropped, changed or replaced by other letters. Table 1 lists some of the most recoded Arabic letters. The following is an example of letter recording for Arabic words.

TABLE I. MOST RECORDED ARABIC LETTERS

Word	Recoding Rule (from→ to)	Word after Recoding
هزئ	ؤ → ئ	هزؤ
	أ → ئ	هزأ
نبئ	أ → ئ	نبأ
خطئ	أ → ئ	خطأ
خسئ	أ → ئ	خسأ
صبئ	أ → ئ	صبأ
سيئ	أ → ئ	سيا
نبء	أ → ء	نبا
دنى	أ → ى	دنا
تؤمن	أ → و	تأمن
ؤمر	أ → و	أمر
راد	و → ا	رود
حيا	ي→ و	حيي

Step6: By the use of stemmers, Words in the collection must be organized into groups, multiple errors are produced and may be used to compare and evaluate stemmers. If the two words belong to the same class of development in meaning and been changed to different origins, this is considered error of under-stemming (i.e. too much of the expressions or terms removed. The stemmer went on correct, if they are changed to the same origin, this will be considered as an over-stemming (i.e. not much of the expressions or words are removed).

IV. RESULTS AND DISCUSSION

Table 2 shows the stemming results of six Arabic sentences varying from sixteen words to two words using the proposed Arabic stemming technique.

TABLE II. ARABIC STEMMING RESULTS

Sentence	Actual Text	Stemmer Text
S1	يختلف اسنان عصر العلم والتقنية الثورة العلمية الخلاقة عن إنسان العصور السابقة في عملية الإبداع والتمكن	خلف سنن عصر علم قنأ ثور علم خلق أنس عصر سبق عمل بدع مكن
S2	طور التعليم في المملكة العربية السعودية مقارنة بدول العالم الاسلامي	طور علم ملك عرب سعد قرن بدل علم سلم
S3	مشاكل العلاقات الدولية وأثرها على مداخيل المواطن العربي	شكل علق دول أثر دخل وطن عرب
S4	طرق التحاق الطالب بالجامعات في اوربا وامريكا	طرق لحق طلب جمع ورب وامريكا
S5	حليل وتصميم انظمة الحاسب الآلي	حال صمم نظم حسب ليي
S6	جامعة نجران	جمع نجر

From the stemming results it is clear that each word returns to its root. The actual sentences and their stemming results are used to see the effect of the Arabic stemming in the famous search engines Google Chrome, Internet Explorer and Mozilla Firefox in terms of total number of the web pages and the Search Time Ratio (STR) which is calculated by the equation 1.

$$STR = Search\ Time * 1000/total\ searched\ page \quad (1)$$

TABLE III. INTERNET EXPLORER SEARCH RESULTS

Internet Explorer				
Actual Sentence			Stemmer Result	
Sentence	Number of searches pages	STR	Number of searches pages	STR
S1	2,290	0.519650655	21,300	0.04835681
S2	171,000	0.004502924	549,000	0.0017122
S3	284,000	0.002640845	616,000	0.00137987
S4	413,000	0.002033898	467,000	0.00152034
S5	467,000	0.00117773	692,000	0.00114162
S6	711,000	0.000675105	86,100,000	5.3426E-06

Table 3 shows the results of the search for actual sentences and their stemming results on the Internet Explorer. From the table the number of searched pages using stemming results is very high compared with the actual sentences results, while the search time ratio for stemming is less than the actual sentences. This means the stemming accelerates and increase the search output in internet Explorer.

Table 4 shows the results of the search for actual sentences and their stemming results on the Google Chrome. From the table the number of searched pages using stemming results is very high compared with the actual sentences results, while the search time ratio for stemming is less than the actual sentences.

TABLE IV. GOOGLE CHROME SEARCH RESULTS

Google Chrome				
Actual Sentence			Stemmer Result	
Sentence	Number of searches pages	STR	Number of searches pages	STR
S1	2,300	0.308696	21,300	0.032394
S2	171,000	0.003626	548,000	0.000949
S3	611,000	0.000753	615,000	0.000813
S4	216,000	0.002546	465,000	0.00086
S5	627,000	0.000829	679,000	0.000766
S6	716,000	0.000656	87,100,000	3.9E-06

Table 5 shows the results of the search for actual sentences and their stemming results on the Mozilla Firefox. From the table the number of searched pages using stemming results is very high compared with the actual sentences results, while the search time ratio for stemming is less than the actual sentences. The search time ratio for Mozilla Firefox is less than Internet Explorer for actual and stemming result even the total search page is comparable, while Mozilla Firefox results is comparable with Google Chrome for the total searched pages and the search time ratio.

TABLE V. MOZILLA FIREFOX SEARCH RESULTS

Mozilla Firefox				
Actual Sentence			Stemmer Result	
Sentence	Number of searches	STR	Number of searches	Search Ratio Time
S1	2,280	0.22807	21,300	0.030986
S2	189,000	0.004074	547,000	0.000951
S3	548,000	0.001113	615,000	0.000732
S4	217,000	0.002442	467,000	0.000835
S5	627,000	0.000797	679,000	0.000736
S6	716,000	0.000712	87,100,000	3.21E-06

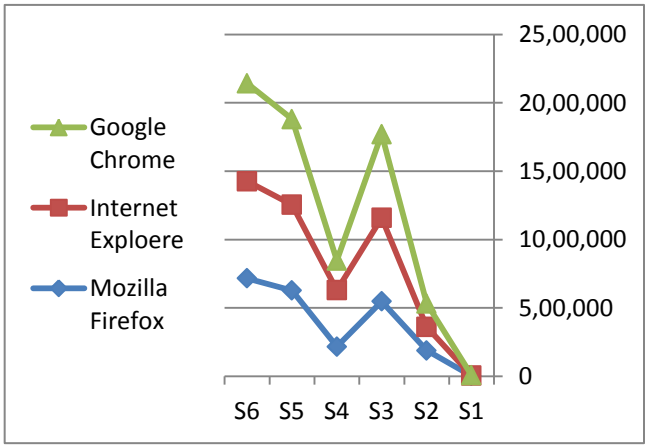

Fig. 1. Actual sentences search results

Figure 1 shows the output of the three search engine for the actual sentences in terms of total searched pages. From the figure it is clear that Google Chrome outperforms Internet Explorer and Mozilla Firefox in the number of searched pages for the actual sentences.

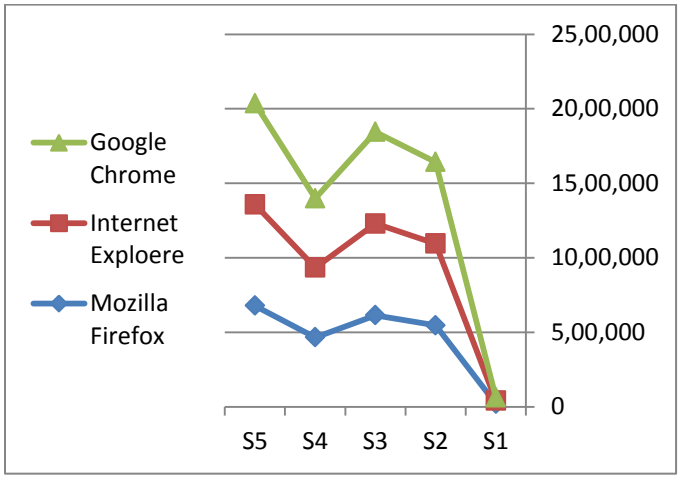

Fig. 2. Stemming results search pages

Figure 2 shows the output of the three search engine for the stemming output. Also Google Chrome outperforms Internet Explorer and Mozilla Firefox in the number of searched pages for the stemming results.

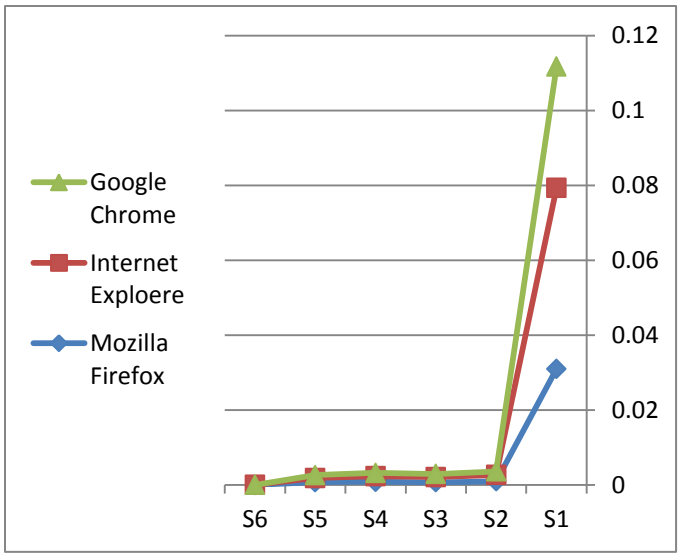

Fig. 3. The str for the stemming results search pages

Figure 3 shows the STR of the three search engine for the stemming results. It is clear Mozilla Firefox takes short time to perform the search output than the Google Chrome and Internet Explorer.

V. CONCLUSION

The fact that there is a lack of usage of Arabic stemming in the search engines gave the underlying rational for conducting the study presented in this paper. It has been shown in the present paper the usage of Arabic stemming increase and accelerate the search engines output. Where Google Chrome outperforms Internet Explorer and Mozilla Fire fox in terms of total number of searched pages.

REFERENCES

[1] Abdelhadi Soudi, Günter Neumann and Antal van den Bosch, "Arabic Computational Morphology: Knowledge-based and Empirical Methods",Arabic Computational Morphology, 3–14. 2007 Springer

[2] Aitao Chen, Fredric Gey, "Building an Arabic Stemmer for Information Retrieval", The eleventh Text REtrieval Conference, TREC 2002, was held at the National Institute of Standards and Technology (NIST) November 19–22, 2002.

[3] Al-Fedaghi, S. S. and Al-Anzi, F. S. A new algorithm to generate Arabic root-pattern forms. In Proceedings of the 11th national computer conference. King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia, pp. 391-400, 1989

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

[5] Aljlayl, M., Beitzel, S., Jensen, E., Chowdhury, A., Holmes, D., Lee, M., Grossman, D., and Frieder, O. IIT at TREC-10. In TREC 2001. Gaithersburg: NIST, 2001.

[6] Al-Shalabi, R. Design and implementation of an Arabic morphological system to support natural language processing. PhD thesis, Computer Science, Illinois Institute of Technology, Chicago, 1996.

[7] Beesley, K. R. Arabic finite-state morphological analysis and generation. In COLING-96: Proceedings of the 16th international conference on computational linguistics, vol. 1, pp. 89-94, 1996.

[8] Bomhard, Allan R.. "Toward Proto-Nostratic: A new approach to the comparison of Proto-Indo-European and Proto-Afroasiatic". Amsterdam: John Benjamins 1984, Publishing Company.

[9] Brent, M. R. Speech segmentation and word discovery: A computational perspective. Trends in Cognitive Science, 3 (8), pp. 294-301, 1999.

[10] Darwish, K., Doermann, D., Jones, R., Oard, D., and Rautiainen, M. TREC-10 experiments at Maryland: CLIR and video. In TREC 2001. Gaithersburg: NIST, 2001

[11] De Roeck, A. N. and Al-Fares, W. A morphologically sensitive clustering algorithm for identifying Arabic roots. In Proceedings ACL-2000. Hong Kong, 2000.

[12] de Marcken, C. Unsupervised language acquisition. PhD thesis, MIT, Cambridge, 1995.

[13] El-Beltagy S., Rafea A.. A FRAMEWORK FOR THE RAPID DEVELOPMENT OF LIST BASED DOMAIN SPECIFIC ARABIC STEMMERS, Proceedings of the Second International Conference on Arabic Language Resources and Tools, 2009

[14] Flenner, G. Ein quantitatives Morphsegmentierungssytem fur Spanische Wortformen. In Computatio linguae II, U. Klenk, Ed. Stuttgart: Steiner Verlag, pp. 31-62, 1994.

[15] Goldsmith, J. Unsupervised learning of the morphology of a natural language. Computational Linguistics, 27 (2), pp. 153-198, 2000.

[16] Goldsmith, J., Higgins, D., and Soglasnova, S. Automatic language-specific stemming in information retrieval. In Cross-language information retrieval and evaluation: Proceedings of the CLEF 2000 workshop, C. Peters, Ed.: Springer Verlag, pp. 273-283, 2001.

[17] Hayder A., Shaikha A., Amna A., Khadija A.,Naila A., Noura A., and Shaikha A., Arabic Light Stemmer: A New Enhanced Approach," in Proceedings of Software Engineering Department, UAE University, Dubai, pp. 1-9, 2005

[18] Kazem Taghva, Rania Elkhoury, Jeffrey Coombs , "Arabic Stemming Without A Root Dictionary", Proceedings of the International Conference on Information Technology, Coding and Computing (ITCC) 2005.

[19] Leah S. Larkey, Lisa Ballesteros , Margaret E. Connell ," Light Stemming for Arabic Information Retrieval", Arabic computational morphology,221-243, 2007 springer

[20] Mohamad Ababneh, Riyad Al-Shalabi, Ghassan Kanaan, and Alaa Al-Nobani "Building an Effective Rule-Based Light Stemmer for Arabic Language to Improve Search Effectiveness" The International Arab Journal of Information Technology, Vol. 9, No. 4, July 2012

[21] Mohammed Naji AL-Kabi, Ronza S. Al- Mustafa, "Arabic Root Based Stemmer", The 2006 International Arab Conference on Information Technology (ACIT'2006)

[22] Mohamed I. Eldesouki, Waleed M. Arafa,Kareem M. Darwish "Stemming techniques of Arabic Language: Comparative Study fromthe Information Retrieval Perspective",The Egyptian Computer Journal , Vol. 36 No. 1, June 2009

[23] Mohammed Aljlayl , Ophir Frieder, "On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach", CIKM'02, November 4-9, 2002, M clean, Virginia, USA.

[24] Nwesri A., S.M.M Tahaghoghi, Falk Scholer, Stemming Arabic Conjunctions and Prepositions, In Mariano Consens and Gonzalo Navarro (eds.), *Lecture Notes in Computer Science - Proceedings of the Twelfth International Symposium on String Processing and Information Retrieval (SPIRE'2005)*, Buenos Aires, Argentina, 3772:206-217, November 2-4,2005.

[25] Paice C.D., "An evaluation method for stemming algorithms". In W.B. Croft and C.J. van Rijsbergen, editors, *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 69-90. Springer-Verlag, July 1994.

[26] Riyad Al-Shalabi, Ghassan Kanaan, Sameh Ghwanmeh, "Stemmer Algorithm for Arabic Words Based on Excessive Letter Locations" 978-1-4244-1841-1/08/$25.00 ©2008 IEEE

[27] S. AI-Fedaghi and F. Al-Anzi, "A new algorithm to generate Arabic root-pattern forms," 11th National Computer Conference, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia, pp. 4-7, 1989.

[28] Shahrul Azman Noah,et el, "soft computing application and indigent systems", second international mutli-conference on artificial intelligence technology, M-CAIT 2013

[29] Youssef Kadri & Jian-Yun Nie, "Effective Stemming for Arabic Information Retrieval", The challenge of Arabic for NLP/MT Conference,2006, The British Computer Society. London, UK.