Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

Improving Information Retrieval Systems' Efficiency

Dr. Loubna Ali
Dept. of Information Technology Engineering
Tartous University Syria

Dr. Shkelqim Hajrulla
Dept. Computer Engineering
Epoka University Albania

***Abstract*— This paper proposes a new stemming algorithm for rooting Arabic words and attempts to solve the polymorphism problem of the word itself by returning it to its root.**
The proposed algorithm will be based on introducing new rules of patterns that increase the efficiency of word identification. Also, this algorithm will contribute to enhancing the efficiency and speed of information retrieval in search engines. Using these rules, he can determine whether a sequence of suffixes is part of the real word or not and remove it.
In this research, a new tool was also developed that allows the user to use any dataset written in Arabic and implement the derivation on it to check the new stem.
To ensure the effectiveness of the proposed algorithm by derivation accuracy test was tested by applying the proposed algorithm to various texts, and then it was compared with Khoja's and a previous algorithms, which were applied on the same data. The results of this research indicated a good improvement in the accuracy of stemming.

I. INTRODUCTION

There is very little research in the area of Arabic computing, and the problem is that non-Arabic researchers do most of the research. Those researchers lack extensive knowledge of the Arabic language itself. These research leads to dead ends or very low and unsatisfactory success rates. One of the most important areas that require intensive research in the specialization of the Arabic language is the field of information retrieval systems.
Information Retrieval (IR) is the art and science of searching for information in documents, for documents themselves or for information describing documents, or searching databases, etc., according to user's need. The main purpose is to retrieve what is useful and to exclude what is not [1].
Stemming techniques are used to increase the effectiveness of information retrieval. Stemmers are essential elements of query systems, Search Engines, classification, and information retrieval systems. Stemming in retrieval systems removes additions from the word to extract its basic formula. The resulting basic formula may not be the lexical root itself, but may be the maximum number of common characters between words [2].
Stemming has two basic types: Root stemming, where the word comes back to its origin by removing all additives, and light stemming which removes fewer additions [2].
The importance of stemming in information retrieval systems has been pointed out by Lennon et al., Which summarizes its usefulness in two points: first, stemming reduces the total number of unique terms and thus reduces the size of the resulting index, as well as that common and similar words in general have a similar meaning. Therefore, retrieval becomes more effective by combining these terms and making them a single term that expresses the common meaning [3].
In practice, the stemming benefit is that the extracted roots are used in text compression, text search, spell checking, dictionary search and text analysis. On the other hand, the removed additions are used to determine the grammatical structure and this is very important for linguists [4].
The impact of stemming on the effectiveness of information Retrieval systems has been the focus of many researches, Of these studies [5] [6] [7].

II. RESEARCH OBJECTIVE

The importance of effective stemming techniques has increased with the huge growth of Arabic content. Many techniques and stemmers have been developed for the Arabic language, but they still suffer from weaknesses and many problems that need to be solved to develop the information retrieval process. There are many challenges that need deep analysis.
Many retrieval applications use the stemmed word to make the search scope wider and larger in meaning, thus ensuring the largest number of relevant and close matches are included in the search results.
When the retrieval system deals with the word without stemming, the word itself will have many different images (morphological variants) and this uses a considerable storage and processing time. Stemming can develop the process of retrieving linked documents, reduce the size of the index, and make query terms more focused on the meaning of the term and its related terms and less focused on the matching of characters.

III. ABOUT ARABIC LANGUAGE

Arabic is one of the most difficult languages - written and spoken. On the other hand, it is one of the most widely spoken languages in the world with more than 400 million speakers as a first language and more than 250 million as a second language. The Arabic language differs from other languages syntactically, morphologically and semantically. The Arabic language belongs to the Semitic languages family, where most of its words are built from the roots by following fixed measurements or known patterns, which makes Arabic

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

different from other languages - being derivative - while most other languages are concatenative, such as English. That is, parts of the word are interrelated sequentially [3] [4].
Arabic language challenges in the field of information retrieval systems can be resumed in the following points [2]:

- Morphological change: is a change in the form of the word itself by adding suffixes or prefixes or any additional letters.
- The phenomenon of stacking, which changes the shape of the letters according to their position in the sentence.
- The abnormal plural of nouns takes a morphological form different from the singular form of the word.
- There is no separation between the word and its prefixes nor its suffixes e.g. the phrase "You asked me" in Arabic is "سألتني".
- The word takes more than one meaning in different contexts although it has the same pronunciation and writing, for example: "ذهب" may come in the sense of yellow metal (gold) in the case of a name, and may come in the sense of (go) in the case of a verb.
- Many words refer to the same meaning, e.g. " ظهر ، بان ، برز" which mean "Appear".
- The meanings of words vary depending on their grammatical status, such as being genitive or subject.
- The Arabic language lacks a sufficient comprehensive source as a kind of database (corpora) for all Arabic vocabularies, such as the English WordNet database.

Stemming Techniques

Light stemmers are based on the process of removing prefixes and suffixes, depending on tables of prefixes and other tables of suffixes that vary from one algorithm to another. This results in correct or incorrect roots, according to the depth of the study and the order of removal of these additions. These algorithms do not look at infixes, or try to match them with patterns. The aim of these algorithms is to reach the stem that is closer to the meaning of the word itself, without trying to find the linguistic root. These algorithms increase the size of the index significantly as a result of the repeated occurrence of the word itself in several close forms [8].

Root-based Technique depends on the patterns of the Arabic language and morphological analysis of the word. It removes all additions related to the word (prefixes, suffixes, infixes), and matches it with the Arabic patterns to produce linguistic root. This process combines largest similar number of words in a single entry in the index, so number of index entries are reduced and that may increase the response speed, but on the other hand it strips the word of its contextual meaning [9].

IV. RELATED WORK

The researchers (Khoja and Garside, 1999) in the study [9] designed a morphological analyzer; based on the rules of morphology and Arabic patterns. It follows this procedure:

1. Remove diacritics representing vowelization.
2. Remove stopwords, punctuation, and numbers.
3. Remove definite article "ال".
4. Remove inseparable conjunction "و".
5. Remove suffixes.
6. Remove prefixes.
7. Match the extracted root against a list of patterns. If a match is found, then extract the characters in the pattern representing the root.
8. Match the extracted root against a list of known "valid" roots.
9. Replace weak letters "ا،و،ي" with "و".
10. Replace all occurrences of hamza "ء،ئ،ؤ" with "أ".

The researchers in the study [11] in 2005 created a stemmer called ISRI that extracts the roots but without using the root dictionary.
The study of [12] in 2014 used a hybrid method in an attempt to overcome the obstacles found in current stemmers. This hybrid method includes two models: light stemming and look in tables.
However, the researcher of [14] in 2016 has proposed an Arabic stemmer that combined the rules of root-based stemmer and light-based stemmer. This research achieved good results, but it applied the rooting algorithm to single words and not to an integrated text.

In a previous research [15] 2022 the researchers proposed new algorithm to improve root-based stemmer and they achieved an important progress. Nevertheless, the study did not take into account the abnormal rules in the Arabic language, such as cracked plurals.

V. THE PROPOSED STEMMER

In this research, a modified algorithm will be proposed to solve the problem of some words that may cause confusion in the rooting process like cracked plurals. The index will be expanded. Where words that may cause a stemming error will be added. Those words will be named special words. So that, the rules of root-based stemmer and light-based stemmer will be combined.

The working mechanism of this research is divided into two phases: The first phase includes preparing the text to be entered into the second phase for applying the stemming algorithm.
The first phase algorithm is shown in the figure (2). In the first time, the text sentences will be divided into words. Then the text will be unified, common and special words will be excluded and the last process before applying stemming algorithm is excluding strange words.

The process of text unification achieve the following :

- Remove stress and diacritics.
- Remove punctuation.
- Remove characters that are not Arabic letters such as numbers.
- Replacing several forms of the letter with one image in general, such as replacing the letter “أ” with its forms “أ”, “إ”, “ا” with “ا” and replacing the “ى” with “ي” at the end of words.

The common words that will be excluded are the separate nominative and accusative pronouns), Imperfect Verbs, Sign names, Conjunctions, Adverbs, etc. Approximately, 210 common words have been listed in this research. Regarding special words, we searched for all cracked plurals in Arabic and found 420 words. We also found about 20 words that represent the plural noun, that is, a word that has a plural

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

origin and cannot be converted into a single word of the same pronunciation.

However, strange words are like months, days, country's names, etc. About 300 strange words have been collected.

The second phase is dedicated to apply the stemming algorithm. In order to get the best results, we have taken into consideration the Khoja algorithm's steps and some ideas that discussed in other algorithms as a basis for this work [15] [17] [18] [19]. In the previous related studies, we observed that removing affixes before performing morphological or grammatical analysis can caused removal of original letters from the word and thus deviates stemming process from its primary purpose. In the previous study [15] grammatical checks prior were performed in the goal to improve stemming results. The stemming algorithm of the previous research was as shown in figure (1)

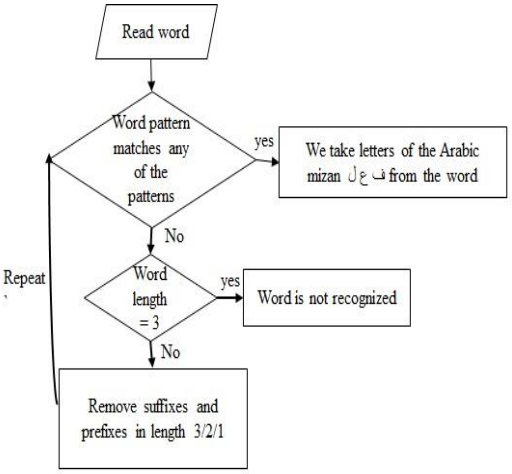

Fig. 1. The stemming algorithm of the previous research

However, that study has faced several problems: the first problem was getting the correct singular for cracked plurals, so we created an index containing these words and the second problem was dealing with words containing three letters. Finally, the problem of processing time, as it was remarkably long when applying the algorithm before modification.

The proposed algorithm for this paper are shown in the figure (2):

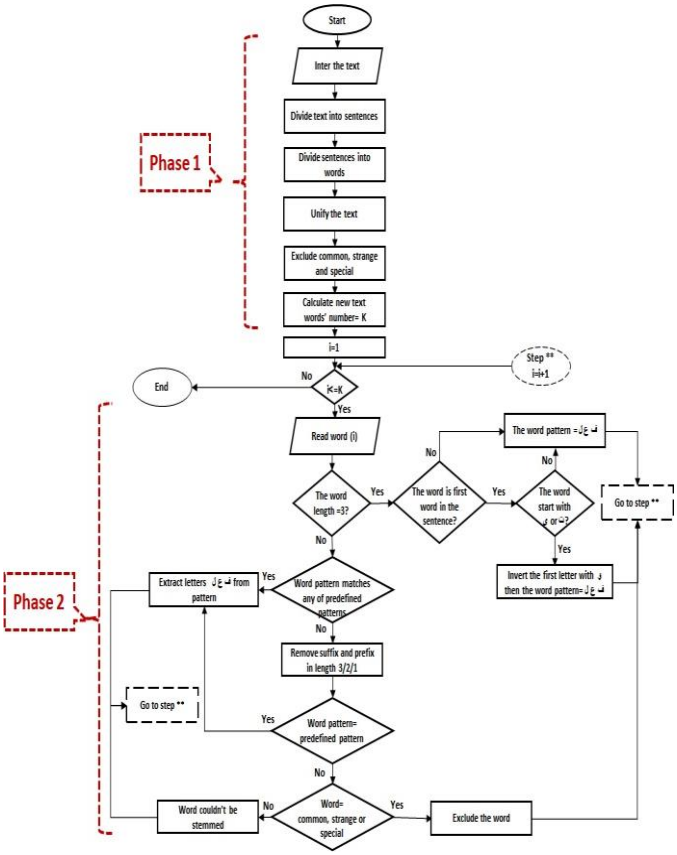

Fig. 2. The proposed algorithm for this paper

The steps of the second phase can be explained as follows:

1. Testing if the word has three letters, if yes, we have now the test if it is a verb or a noun. When the word is in the first of the sentence then the word will be considered as a verb otherwise the word will be considered as a noun. If the word in the first of the sentence and start with "ت" or "ي" it will be consider as a verb in present that derived from a past verb with a vowel in the middle, which is the letter "و", for example "يجب، وجب"، "يهب، وهب"، "يثب، وثب", etc.

We know that the method adopted in this research will not produce 100% correct results, but it will improve the results of the previous search, which completely excluded working with three-letter words.

2. Testing if the word is matched with one of the patterns, and if it does not match any pattern in the patterns' group shown in Table 1, the process of removing affixes will start.

Length	Patterns
4	فاعل-افعل-فعّل-تفعل-فعال-مفعل-فعول-فعيل-فعلى-فعلة
5	افتعل-انفعل-تفاعل-تفعيل-افعال-مفاعل-متفعل-منفعل-مفتعل-مفعول-مفعال-فعلان-فعلاء-فواعل-افاعل-يفتعل-تفتعل-فاعول—فعائل-تفعلل-افعلل-مفعلل-فعالل
6	استفعل-انفعال-افتعال-افعلال-متفاعل-مستفعل-مفاعيل-يستفعل-افعوعل-متفعلل
7	استفعال

a. Groups of Arabic Patterns used in Algorithm

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

3. Removing prefixes and suffixes in length of three then two then one as shown in Table 2 starts respectively.
4. Retest matching with patterns.
5. Restart with new word until finishing all words in the taxt.

Affixes		
prefixes	***P1***	ل- ب – ف – س – و – ي – ت – ن – ا
	P2	ال – لل – فل – ول – وب
	P3	ولل – وال – كال – بال – فال
suffixes	***S1***	ة – ه – ي – ك – ت – ا – ن
	S2	ون – ات – ان– ين – تن – كم – هن – نا – يا – ها – تم – كن – ني – وا– ما – هم
	S3	تمل – هما – تان – تين – كما

b. Groups of Arabic Affixes used in Algorithm

VI. RESULTS

In the goal of realizing the algorithm, Java programming language was used on several Arabic paragraphs. In order to diversify the tested words, several topics have been selected such as topics that talk about economics, sports, education and several more.

Those paragraph have been chosen from data sample called OSAC (Open Source Arabic Corpora) [14], which is a database containing articles in Arabic in various scientific, political, sports and economic fields for the purpose of this research three paragraph were chosen containing 1558 words.
To clarify the results presented by this research regarding the validity of the rooting process, we will present the example shown in Table 5. We used a list of terms as a sample, which is the same sample used in [9,13], then the derivation process was tested using the derivation algorithm (Khoja), the previous and the algorithm proposed in this research.

The results after applying the proposed algorithm were compared with Khoja and previous algorithms results. The following table shows some examples of those results. We can consider that the word stemming using proposed algorithm lead to better results than Khoja algorithm and the previous algorithm (fault stemming words are red and underlined)

The word	**khoja**	**Previous**	**Proposed**
فلتستعجله	فلتستعجله	عجل	عجل
المكتبات	كبي	كتب	كتب
كامل	كمل	كمل	كمل
منوعات	وعي	نوع	نوع
يمدّ	مدد	مدد	مدد
تستغرق	تستغرق	غرق	غرق
سباق	سبق	سبق	سبق
الخيل	خيل	خيل	خيل
العاشر	عشر	عشر	عشر
كانون	كان	كانون	كانون
يفي	في	يفي	وفي
يرى	ري	يري	وري
متفاهمون	فهم	فهم	فهم
المنظمات	ظما	نظم	نظم
دعيت	دعا	دعا	دعا
رأيت	رأى	رأى	رأى
يجب	جب	يجب	وجب

c. A comparison between stemmers

Other results that we obtained are shown in the following table:

criterion	**Khoja**	**Previous algorithm**	**Proposed algorithm**
Total number of words	1558	1558	1558
Number of stop-words	354	480	480
Number of stop-words with أو , و	410	536	536
Number of strange words	-	79	79
Number of special words	-	-	23
Number of both stop-words, special and strange words	410	615	638
Number of stemmed words	1148	943	920
Deduction percentage	-26.3%	-39.5%	-40.9%
Stemming time	42 sec	35 sec	23 sec
Number of correctly stemmed words	640	709	867
Stemming accuracy	56%	75%	94%

d. A comparison between previous algorithm, Khoja and proposed algorithm.

The results show that Khoja could not always arrive to extract the correct root of the word. In many cases, it produces an entirely new word, often wrong and does not exist in Arabic language. The previous stemmer algorithm succeeds to extract the right root of the word and effectively removes all affixes and do not remove any original characters from the word but it does not treat the three letters words correctly. While the proposed stemming algorithm deal more efficiently with the three letters words and exclude special words from text.

In order to analyze the results we obtained, we will start with the percentage of word reduction that the first stage will lead to for each of the algorithms. So that the deduction percentage equation can be used:

$$Percentage\ of\ deduction = \left[\frac{SW - W}{W}\right] * 100\% \qquad (1)$$

SW: number of words after removing stop-words and common, special and strange words
W: total number of words

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

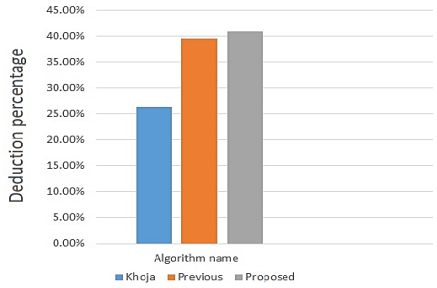

Fig. 3. deduction percentage.

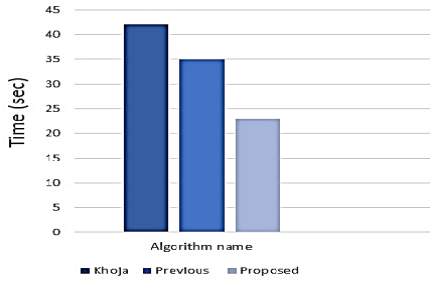

Fig. 4. Processing time.

The greater the deduction percentage in absolute terms, the better it is reflected in the information retrieval system, as the volume of processed information decreases, and thus leads to an increase in the speed of obtaining the required results. As Figure 3,4 shows that the proposed algorithm gives better results in deduction percentage and processing time.

In order to evaluate the performance of the proposed stemmer. Accuracy of stemming is defined as:

$$Accuracy = \frac{correctly\ stemmed\ words}{total\ stemmed\ words} \quad (2)$$

We got a total of 867 correctly stemmed words out of 920 words that entered the stemming process according to the proposed algorithm. While using previous algorithm we got about 709 correctly stemmed words out of 943 when applied to same data set. However, Khoja gave back about 640 correctly stemmed words out of 1148 when applied to same data set. As consequence, the different accuracies are displayed in figure (5).

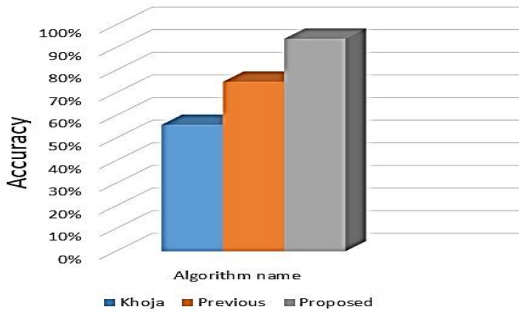

Fig. 5. the accuracies of the three algorithms.

Corrected stemmed words factor (CSWF): It is the percentage of correctly stemmed words relative to the number of words stemmed by the stemming algorithm. The higher the ratio, the higher the accuracy of the stemmer. We obtained this ratio by applying the stemmer to a data sample: OSAC (Open Source Arabic Corpora) [14], a database containing articles in Arabic in various scientific, political, sports and economic fields, the result was 94%, which is higher than the percentage obtained by Khoja algorithm which is 56% [16], and in previous algorithm which was 75% [15].

ACKNOWLEDGMENT

This paper has presented a way of stemming Arabic language based on morphological rules. New algorithm based on Arabic grammar and patterns was been developed, this algorithm improved average stemming accuracy and thus increased the efficiency of word identification. The purpose was achieved by checking Arabic word's pattern before removing affixes is very important criteria to a successful stemming. Then removing strange and stop-words reduces time and size of storage required by the processed data and finally using these rules, we can determine whether the sequence of affixes is part of the original word or not and thus solve the problem of ambiguity.
Our study can help in development of many systems such as automatic dictionaries, document classification, automatic translation as well as information retrieval systems.

REFERENCES

[1] L. Ali, M. Jaber, S. Chaari, F. Biennier, "Context-aware infrastructure to support distributed industrial services, Emerging Technologies and Factory Automation ," ETFA. IEEE Conference , 2007, pp. 716-719.

[2] A. Al-Omari, B. Abuata, " Arabic Light Stemmer (ARS)," Journal of Engineering Scienceand Technology, vol. 9, no. 6, 2014, pp. 702-716.

[3] L. Anghel, R. Velazco, S. Saleh, S. Deswaertes, A. El Moucary, "Preliminary validation of an approach dealing with processor obso-lescence ," 18th IEEE Int. Symp. Defect and Fault Tolerance in VLSI Systems, vol. 6, no. 3, 2003, pp. 493.

[4] H. Ballaoui, H. Lahmar, N. Labani, "Information Retrieval in Arabic Language: an Approach Based on the Defining Clauses," International Review on Computers and Software , vol. 11, no. 6, 2016.

[5] K. Darwish, D. Orad, "CLIR experiments at Maryland for TREC-2002: Evidence combination for Arabic-English retrieval," The Eleventh Text REtrieval Conference TREC-11 conference, vol. 3, no.2, 2002, pp. 703-710.

[6] K. Darwish, W. Magdy, Arabic Information Retrieval, now Publishers Inc, 2rd edition, United States, 2013.

[7] A. Gowder, I. Almerhag, A. Ennakoa, "Arabic Broken Plural Recognition Using A Machine Translation Technique," *Journal of Information Science*, vol. 2, no. 5, 2008, pp. 300-312.

[8] M. Ibrahim, S. Rizvi, "a metasystem base representation of the standard system data dictionary," second national conference (MATEIT-2008) mathematical techniques,vol. 2, no. 3, 2008.

[9] I. Jomaa, Strucutres of Arabic Language, Dar Al Fahed , Saudi Arabia, 2nd edition, 2006.

[10] S. Khoja, R. Garside, "Stemming Arabic text," Computing Department, Lancaster University, vol 1,1999, pp. 200-210.

[11] A. Kreaa, A. Ahmad, K. Kabalan, "Arabic Words Stemming Approach Using Arabic Wordnet," International Journal of Data Mining & Knowledge Management Process, vol. 4, no. 6, 2014.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 11, November-2022

[12] L. Larkey, L. Ballesteros, M. Connell, "Improving stemming for Arabic information retrieval: Light stemming and co-occurrence analysis," Proceedings of the 25th International Conference on Research and Development in Information Retrieval (SIGIR), vol 3, no.6, 2002, pp. 275-282.

[13] M. Lennon, D. Pierce, B. Tarry, P. Willett, "An evaluation of some conflation algorithms for information retrieval," *Journal of Information Science*, vol. 3, 1981, pp. 177–183.

[14] R. Mohammed, " New Arabic Stemming based on Arabic Patterns," Iraqi Journal of Science, vol. 57, no. 6, 2016, pp. 2324-2330.

[15] M. Mohammed, L. Ali, M. Ibraheem, "Improve Arabic Stemming in Information Retrieval Systems" Master thesis, Tartous University, 2022.

[16] S. Motaz, A. Wesam, "OSAC: Open Source Arabic Corpora," the International Conference on Information Technology: Coding and Computing, vol. 5, no. 2, 2010, pp. 300-320.

[17] A. Nehar, D. Ziadi, H. Cherroun, Y. Guellouma, "An Efficient Stemming for Arabic Text Classification," Innovations in Information technology (IIT), vol.5, no. 3, 2012, pp. 328-332.

[18] S. Sirsat, V.Chavan, H. Mahalle,"Strength and accuracy Analysis of Affix Removal Stemming Algorithms," (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 4, no. 2, 2013, pp. 265-269.

[19] K. Taghva, R. Elkhoury, J. Coombs, "Arabic stemming without a root dictionary," In Proceedings of the Internaional Conference on Information Technology: Coding and Computing, vol. 1,2014, pp. 152-157.