Conditional Arabic Light Stemmer: CondLight

Yaser Al-Lahham, Khawlah Matarneh, and Mohammad Hassan
Computer Science Department, Zarqa University, Jordan

Abstract: *Arabic language has a complex morphological structure, which makes it hard to select index terms for an IR system. The complexity of the Arabic morphology caused by multimode terms, using diacritics, letters have different forms according to its location in the word and affixes can be added at all locations in a word. Several methods were proposed to overcome these problems; such as root extraction and light stemming. Light stemming show better retrieval efficiency, Light10 is the best stemmer among a series of light stemmers, it simply removes suffixes and prefixes if it is listed in a predefined table. Light10 has no restrictions on the affixes, so it is possible to have two different terms having the same token while they have different meanings. This paper proposes CondLight stemmer which adds new prefixes and suffixes to the table of Light10, and imposes a set of conditions on removing these affixes. The implementation and testing of the proposed method show that CondLight gains 38% precision, while Light10 stemmer gains average precision of 36.7%. Moreover CondLight show better average precision either when imposing all conditions or part of them.*

Keywords: *Arabic IR,light stemming,morphological analysis, affixes' removal, term selection, Arabic document indexing.*

Received February 14, 2018; accepted April 18, 2018

1. Introduction

Arabic language has complex morphological and orthographic forms making it difficult to develop efficient Arabic information retrieval systems. Moreover the written Arabic letters have different forms depending on its location in the word, for example the letter (س) has the following forms: ـس at the end of the word, ـســ in the middle of the word, and ســ at the beginning of the word. For information retrieval, this complex nature results a mismatch between the form of a word in a query and the forms found in documents relevant to the query [24]. On the other hand, words preceded by some preposition (ب ,ك ,ل) are agglutinated to them, making no distinction between the word and the preposition. In fact, many affixes, other than prepositions, have no distinction between the words they agglutinated to; for example: one letter prefixes indicate the gender of the person in the present tense verbs; such as يكتب and تكتب.

In written Arabic, the vowels (diacritics) are omitted, so several words of different meaning will have the same shape. For example, the word (حسب) could has several meanings; it means: (calculate), (think that), or (according to). This ambiguity makes a crucial problem in information retrieval, since an Arabic word can have several meanings.

Another problem of the Arabic language, is the plural form of the irregular nouns, also called broken plural, such words have different forms than its singular form. A solution to this problem is to have a dictionary of such words, which is not always available [16].

Solutions to these problems follow different approaches, which in general aimed to truncate affixes out from words so as to map them to a same token. The most common methods to solve this problem used root extraction, since Arabic words are derived from a limited number of roots, mainly of three letters, and less number of words has roots of four and five letters. Words can be derived by applying some morphological rules. Root extraction assumes that words of a same root have similar semantic, but in many cases it is not, as indicated in [22].

Another approach that shows better performance than root extraction is stemming, and the light stemming in particular has better retrieval precision than root extraction [20].

In Arabic, many stemmers were developed, but the most effective stemmers for information retrieval are the light stemmers, and the best empirical results were recorded by the Light10 stemmer [19].Light10 stemmer is a table-driven stemmer; i.e., it has a list of affixes such that they are removed from any word that its affix matches any table entry or entries. These affixes are removed without aware of the word status; i.e., it could be a part of the word for example.

The main objective of this paper is to improve Arabic information retrieval by improving the effectiveness of the Light10 stemmer by adding extra prefixes and suffixes to the table, and imposing some rules that satisfy the nature of the morphology of the Arabic language, according to detailed study of Arabic words as illustrated in [3].

These rules attempt to overcome some problems that light stemmers cause; such as over-stemming, miss-stemming and under-stemming. These drawbacks decrease the effectiveness of stemming algorithms, however these problems are mutually correlated, that means an attempt to reduce the effect of one type may has negative consequence to the other [23].

The rest of the paper is organized as follows:

- Section 2 presents a review of the literature related to Arabic term selection; light stemming in particular.
- Section 3 introduces the proposed enhancement over Light10 stemmer.
- Section 4 includes the implementation and evaluation details.
- Finally section 5 concludes the results of this research.

2. Related Work

Arabic language term selection includes four types of stemming techniques, namely: manually constructed dictionaries, root-based stemming (morphological analyzers), statistical stemmers and light stemmers. In fact each of these techniques represents a level in a scale of analysis.

Root-based stemmers use morphological analysis to extract the root of a given Arabic word. Among the most popular root-based stemmers; Khoja *et al.* [17] removes suffixes, infixes, and prefixes. It uses pattern matching to extract the roots of words. Khoja *et al.* [17] suffered from some drawbacks especially with some words that represent 'broken plurals'. Khoja *et al.* [17] algorithm was improved by Taghva *et al*. [25], by eliminating the need for a dictionary to support their root extraction, thus eliminating the intensive maintenance and system requirements associated with a dictionary of the entire Arabic language, and Al-Kabi [5] improved this stemmer by adding extra patterns and rules. Recently, pattern-based stemmers root extraction methods are developed without using a dictionary, [6, 22] are examples.

This paper focuses on light stemming, so this category will be presented in more detail in the rest of this section. The light stemming approach, or affix removal, is a process of stripping off a small set of prefixes and/or suffixes, without trying to deal with infixes, or recognize patterns and find roots [19].Rule based stemmers apply conditions to differentiate between the added affixes and the original part of the word, for example; Ababneh *et al.* [1] proposed a rule-based light stemmer, by defining a set of possible prefix and suffix. They attached possible antefixes and prefixes, and possible suffixes and postfixes to the suffix list. An Arabic word pattern is used to determine whether the affix of a word is an original part of the word or not. In case the word does not match any of the predefined Arabic patterns, they tested the relationship between the suffix and prefix, where a suffix or a prefix of a word is an original part of it if it didn't appear in the predefined list.

Regarding statistical methods, Darwish and Ali [10] presented a method for generating morphologically related tokens from Wikipedia hypertext to page title pairs, by altering the spellings of transliterated named entities. Boudlal *et al.* [8] presented a morphological analyzer for unvoweled Arabic sentences, such that it gives multiple roots to a single word. The right root is selected among the resulted roots according to the context in which it presents. Hidden Markov model is used to discover the root given its context.

Hybrid methods combine stemming technique with some other techniques. For example Mustafa [21] used n-gram string matching presented in [18], with a light stemming method. Alhanini and Aziz [2] used light stemming and dictionary-based stemming. Hadni *et al.* [13] combined Khoja Stemmer, Light Stemmer and n-gram techniques. Khedr *et al.* [15] included some rules to both stemmers of EL-Beltagy and Rafea [11] and Light10 stemmer.

Affix removal light stemming algorithms; for example [4, 9] introduced a very close set of affixes' tables with certain conditions to strip prefixes and suffixes from words. Each of them define a set of rules that applied to words in order to remove prefixes or suffixes; such as removing definite articles for prefixes, or removing pronouns for suffixes.

Light10 proposed by Larkey *et al.* [19, 20] as a developed version of a series of Arabic light stemmers. Light10 is the most effective of this series. They concluded that light stemming improves retrieval without providing correct morphological roots. Since Light10 is the basis of the proposed CondLight stemmer, so it is worth to describe it in detail.

Light10 applies the following procedure:

- Remove the letter و (“and”) if the remainder of the word length exceeds, or equal to three characters. Even the removal of the letter 'و' is helpful; it may cause some problems, since many Arabic words begin with this letter as an original part of it.
- Remove any of the definite articles, if two or more characters left, and finally,
- Remove affixes from a word that match any item in a predefined list, if the length of the word left is at least two characters. The list of prefixes is: و, لل, فال, ال and the list of suffixes is: ي, ة, ه, يه,ية, وال, بال, كال, ها,ان,ات,ون,ين.

As a consequence, most of the work in the literature focused on improving the accuracy of the morphological analyzer; either it is a root extraction based, a heavy stemmer, or a light stemmer. Some research efforts tested these methods against retrieval on a real world test collection; such as Larkey *et al.* [19, 20].

3. Conditional Light Stemmer: Condlight

The methodology of the proposal in this paper is to investigate the morphological aspects of Arabic language, determine the cases in which affixes could be removed, build conditions according to this investigation, and finally adding extra conditional

affixes to the table of the Light10, as will be explained in this section.

Light stemming is used to unify Arabic terms in a simple way without dealing with the complex morphological nature of these words. The unified terms (or tokens) can be used to index Arabic documents in an IR system, for example: different documents that use different forms of a term that has the same meaning will be indexed, even if they do not contain that form. Moreover these documents will be returned as a result of a user need, even they do not include the form of a term entered by the user. Table1 presents an example of such type of retrieval.

Table 1. IR Indexing without, and with stemming.

Without Stemming:					
المعلم	User query				
Original term	والمعلم	فالمعلم	كالمعلم	المعلمون	المعلمات
Index term	والمعلم	فالمعلم	كالمعلم	المعلمون	المعلمات
With stemming:					
المعلم	User query				
Original term	والمعلم	فالمعلم	كالمعلم	المعلمون	المعلمات
Index term	معلم	معلم	معلم	معلم	معلم

This paper proposes adding extra prefixes to the list that has been used in Light10 stemmer, which are: (بـ ,تـ ,ف ,لـ ,سـ ,يـ) and new suffixes(هن , هم, وا). Light10 is used as a basis for this new proposed stemmer since it proves effective in Arabic IR, and it is recommended by recent research efforts such as in [14]. In order to remove these prefixes and suffixes, a set of conditions should be satisfied; these conditions are derived from the morphological properties of the Arabic language, and could be stated as follows:

- *Condition* 1: To remove the suffixes:(هن , هم) from the end of a token, it should not begin with one of those prefixes (وال, كالـ , فالـ , الـ ,لل ,بال), since in these cases, these suffixes are original parts of words, for example(التقاهم), begins with prefix (الـ) and ends with the suffix (هم), which is an original part of this word, so there is no need to remove it in this case. In another example (كتبهم), the suffix (هم) is not original part of the word, so it must be removed.
- *Condition* 2: To remove the prefix (يـ) from a word, it must ends with one of these suffixes (وا, ة, هان, ون, ها), as in the words: (يفعلها, يفعلون, يفعلان, يفعله, يفعله, يفعلوا), these words are present verbs, so the prefix (يـ) is not original part of token, so it must be removed.
- *Condition* 3: To remove the prefix (تـ) from a word, it must ends with one of these suffixes (هان,ين,وا ة, ون, ها, ت), as in the words: تفعلان , تفعلون , تفعلها ،تفعلوا, تفعلهتفعلين ,تفعلة,, since these words are present verbs, for example: (تواجهها), the prefix (تـ)is not an original part of word it just used to refer present verb for a female, so it must be removed.
- *Condition* 4: To remove the prefix (بـ) from a word it must ends with one of following suffixes (هن , هم, ه بكتبها, بكتابين , بكتابات , ها, ، ين, ات , ان), as in the words: (بكتابان, بكتبة, بكتبه ,بكتبهم ,بكتبهن) since these words are nouns, and this prefix is not an original part of them, for example : (بحادثة),the prefix (بـ) is not an original part of the word, it is a preposition that means (with),so it must be removed.
- *Condition* 5: To remove the prefix (لـ) from a word it must ends with one of following set of suffixes (هم , لكتبها , لكتابين هاين,ات,ان, ة,هن), as in the words: (لكتابات , لكتابان, لكتبة, لكتبه ,لكتبهم ,لكتبهن since these words are nouns, and this prefix is not an original part of them, for example:(لعبته)here the prefix (لـ) is not an original part of this word ,it is a preposition that means (to),so it must be removed.
- *Condition* 6: To remove the prefix (فـ) from a word it must ends with one of following set of suffixes: (فكتبها , فكتابين هاين,ات,ان,ة,هم , هن), as in the words: فكتابات , فكتابان, فكتبة, فكتبه ,فكتبهم ,فكتبهن since these words are nouns, and this prefix is not an original part of them, for example(فعلمة), here the prefix (فـ) is not an original part of this word, it is a preposition that means (then),so it must be removed.
- *Condition* 7: To remove the prefix (سـ) from a word, it must be followed by either(تـ) or (يـ), where condition 2 is then applied, for example in the word (ستدرسها), the prefix (سـ) followed by (تـ) and it means (will),so it is not an original part of this word, so it must be removed.
- *Condition* 8: To remove the prefix (وا)from a word and delete (ا) from its beginning, it must begin with (ا), (يـ) or (تـ), and ends with (وا), since this token is a verb that denotes a plural, for example (اتبعوا)the prefix (ا)is not an original part of this word, it is an imperative verb, and the suffix (وا) also is not an original part of token, so it must be removed.

All of these conditions that control the removal of affixes are deducted from the morphological properties of the Arabic language; such as nouns and verbs forms. The possible affixes added to nouns and verbs are derived from [3], where the nouns and the verbs could be recognized as follows:

1. Present tense has the following cases:
   a. Present tense could be started by: أ, ن, ي, or ت, it could also be recognized by the letter س that used to denote future events. Present usually follow two words: سوف and لن.
   b. The possible suffixes of a present tense are: corroborative nun, such as "يجلسن", double pronoun "ضمير الاثنين", for example; "تقضيان". For female a present verb could has the suffixes: "ين", and "ها", and "وا" and "ه" for singular and plural male respectively.

Condition (2) is imposed according to the present verb possible suffixes and prefixes, where the two prefixes "ت" and "ي" are matched with the possible suffixes as explained above. Condition (4) is imposed since the prefix that denotes future, which is "س", matched with

the two suffixes: "و" and "ت" in the present tense.

2. Imperative verbs: Can be recognized by the following directives: it could starts by "أ" such as "أكتب" means "write", or "ل" as in "لتعمل" means "to do". It could be terminated by "ن" for female plural, by "وا" for male plural, by "ا" for couple entities, and "ي" for female singular. These imperative verb properties related the removal of the prefix "ا" by the suffix "وا", as in condition (8).
3. Nouns: a noun can have different forms; it could be: a name for an entity, comparative (or preeminence), augmented, and un-augmented. A noun could be started by "بـ" and "ف" proposition letters to give causality, for example "فتنجحادرس" means "study to succeed", and "ل" as in "لطيب قلبك" means "for your kind hurt", "ل" could give possessive meaning to a noun. A noun could have the following affixes: "ان" or "ين" for coupled entities. Plural male and has the suffixes: "ون", "ين", and "ات" for female, apparent connected third pronoun; such as "هم", "هن", "هـ", and "ها" are also possible suffixes of nouns.

Consequently, conditions: (4, 5) and (6) where set to connect these possible prefixes of a noun to its possible suffixes.

The list of the affixes used by CondLight, compared to the list used by light stemmer variants, is presented in Table 2.

Table 2. List of removed affixes for different light stemmers.

	Remove prefixes	Remove Suffixes
Light1	ال ,وال,بال, كال,فال	None
Light2	ال ,وال,بال, كال,فال ,و	None
Light3	"	ة,ه
Light8	"	ها,ان,ات,ون,ين,يه,ية,ة,ه,ي
Light10	ال ,وال,بال, كال,فال ,لل ,و	"
CondLight	ال ,وال,بال, كال,فال ,لل ,و بـ ,تـ , لـ , يـ , فـ ,سـ	ها,ان,ات,ون,ين,يه,ية,ة,ه,ي هم , هن, وا

In the next section, the proposed affix removal scheme will be applied on a collection of documents, and will be tested against Arabic documents' retrieval.

4. Implementation and Evaluation

The proposed light stemmer CondLightis applied on the TREC2002 collection [12], which contains 383872 documents drawn from the archive of the Associated France Press (AFP) news wires. These documents are organized in files, where each file contains a day news from the year 1994 to 2000. The collection –also- has a set of 75 user needs called 'topics' numbered from AR1 to AR75, with relevance judgment. Each topic has three parts: header, description, and narrative. In this implementation, a query is made up of the header and the description of each topic.

The documents are indexed in an inverted table such that stop words and punctuations are removed. Terms are normalized such that all (أ آ إ) are replaced by (ا), and (ة) is replaced by (ه).

The same Java class that used to apply the light10 stemmer is implemented and used to apply CondLight stemmer, such that the proposed modifications and conditions are added to generate the documents' index. Table 3 shows a portion of statistical data resulted from the application of the implemented stemmers.

Statistical data of both Light10 and CondLight better enhance retrieval over using no stemming (All Terms), at the same time both stemmers used less number of tokens.

Table 3. Statistical Data of both Light10 and CondLight compared to the baseline (All terms).

	Light10	CondLight	All terms (no stemming)
Number of tokens	284049	258480	473025
Number of postings✦	36945596	36794313	59278987
Average Precision	0.367	0.384	0.254
✦Number of occurrences of all terms in all documents			

CondLight uses only 55% of the total number of tokens and 62% of postings as well. In spite of that CondLight indexed less number of tokens than the number of tokens indexed by light10 (91% of that of Light10); they have close number of postings, that means CondLight discards some terms with no significance that Light10 considered.

Experimental results show that the conditional light stemmer gains about 5% enhancement of retrieval over the original Light10 stemmer. The enhancement occurs mainly at lower recall levels (as shown in Figure 1), which satisfies users' needs by retrieving more relevant documents at the first page; for example of a web search engine.

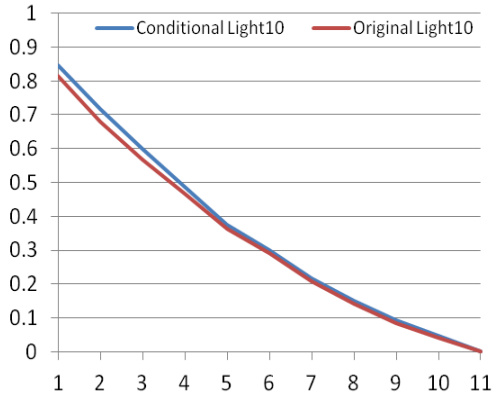

Figure 1. 11-recall levels precision-recall curve for Light10 and CondLight stemmers.

More experiments were applied, such that each time a single condition is applied separately on Light10, the results show that some conditions are more effective than others; as shown in Figure 2. In this figure method number 6 has the highest average precision, this method represents the application of both conditions (6) and (7), and method number 7 represents applying condition (8). Condition (3) has the lowest average precision. As a consequence, applying conditions 6 and 7 could be sufficient to enhance Light10; the same result could be noted for both conditions (1 and 2).

Results enhancement of the proposed technique could be justified as some terms (of different meanings) will have the same morphological shape after affixes removal, making a system retrieves more irrelevant documents; i.e. increasing the false positive retrieval, which yields a lower precision.

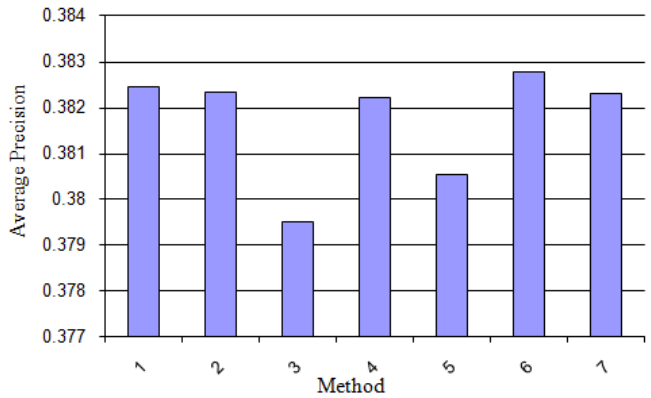

Figure 2. Average precision for each condition applied alone.

When applying the proposed conditions, this situation becomes less probable, hence the chance to have true negative retrieval becomes lower, and consequently enhancing the precision ratio. Moreover, it is found that some words having different shapes (with same meaning) in the query are mapped by Light10 to different tokens, which gives lower weight to both of them, as a consequence, giving lower similarity of the query to some relevant document that may not retrieved, which decrease the precision. These words are mapped to the same token in CondLight (as shown in Table 4),which gives higher weight to the same token, and gives higher similarity to more relevant documents, which enhances the retrieval.

Table 4. A List of terms resulted after applying both light10 and CondLight.

	After applying the stemmer:	
Original Word	**Light 10**	**CondLight**
يدرسها	يدرس	درس
بدخوله	بدخول	دخول
يبذلها	يبذل	بذل
تطورت	تطورت	طور
لمهرجانات	لمهرجان	مهرجان
باضافة	باضاف	اضاف
يلعبة	يلعب	لعب
بحادثة	بحادث	حادث
تسببها	تسبب	سبب
تمثلت	تمثل	مثل
بهجمات	بهجم	هجم
تعاملت	تعامل	عامل

5. Conclusions

In this paper, a conditional light stemmer (CondLight) is proposed as an extension to the Light10 stemmer. The extension includes adding a set of new affixes to be removed if they satisfied one or more of a set of proposed conditions; these conditions are derived from the morphological nature of Arabic words. The application of the proposed light stemmer shows that adding some conditions to the extended light stemmer enhances the retrieval especially at lower recall levels. Moreover, it is found that the application of a single condition will enhance the retrieval, where some conditions (such as conditions 1, 2, 4, 6, 7, and 8) enhance the retrieval when applied separately. As a future work the proposed conditions will be extensively tested on other test collections, and on different light stemmers.

References

[1] Ababneh M., Al-Shalabi R., Kanaan G., and Al-Nobani A., "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, pp. 368-372, 2012.

[2] Alhanini Y. and Aziz M., "The Enhancement Of Arabic Stemming By Using Light Stemming and Dictionary-Based Stemming," *Journal of Software Engineering and Applications*, vol. 4, no. 9, pp. 522-526, 2011.

[3] Al-Hamlawi A., شذا العرف في فن الصرف, Dar El Fekr El Araby, 1999.

[4] Aljlayl M. and Frieder O., "On Arabic Search: Improving The Retrieval Effectiveness Via A Light Stemming Approach," *in Proceedings of the 11th International Conference on Information and Knowledge Management*, Virginia, pp. 340-347, 2002.

[5] Al-Kabi M., "Towards Improving Khoja Rule-Based Arabic Stemmer," *in Proceedings of Applied Electrical Engineering and Computing Technologies*, Amman, pp. 1-6, 2013.

[6] Al-Kabi M., Kazakzeh S., Abu Ata B., Al-Rababah S., and Alsmadi I., "A Novel Root Based Arabic Stemmer," *Journal Of King Saud University-Computer and Information Sciences*, vol. 27, no. 2, pp. 94-103, 2005.

[7] Boudchiche M., Mazroui A., Bebah M., Lakhouaja A., and Boudlal A., "AlkhalilMorpho Sys: A Robust Arabic Morpho-Syntactic Analyzer," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 2, pp. 141-146, 2017.

[8] Boudlal A., Belahbib R., Lakhouaja A., Mazroui A., Meziane A., and Bebah M., "A Markovian Approach For Arabic Root Extraction," *The International Arab Journal Of Information Technology*, vol. 8, no. 1, pp. 91-98, 2011.

[9] Chen A. and Gey F., "Building an Arabic Stemmer for Information Retrieval," *in Proceedings of Text Retrieval Conference*, pp. 631-639, 2002.

[10] Darwish K. and Ali A., "Arabic Retrieval Revisited: Morphological Hole Filling," *in Proceedings of the 50th Annual Meeting of the*

Association for Computational Linguistics: Short Papers-Volume 2, Jeju Island, pp. 218-222, 2012.

[11] El-Beltagy S. and Rafea A., "An Accuracy Enhanced Light Stemmer for Arabic Text," *ACM Transactions on Speech and Language Processing*, vol. 7, no. 2, 2011.

[12] Gey F. and Oard D., "The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic Using English, French Or Arabic Queries," *in Proceedings of The 10th Text Retrieval Conference*, pp. 16-23, 2001.

[13] Hadni M., Lachkar A., and Ouati k., "A New and Efficient Stemming Technique for Arabic Text Categorization," *in proceedings of International Conference on Multimedia Computing and Systems*, Tangier, pp. 791-796, 2012.

[14] Jaafar Y., Namely D., Bouzoubaa K., and Yousfi A., "Enhancing Arabic Stemming Process Using Resources and Benchmarking Tools," *Journal of King Saud University- Computer and Information Sciences*, vol. 29, no. 2, pp. 164-170, 2017.

[15] Khedr S., Sayed D., and Hanafy A., "Arabic Light Stemmer for Better Search Accuracy," *International Journal of Cognitive and Language Sciences*, vol. 10, no. 11, pp. 3587-3595, 2016.

[16] Kadri Y. and Nie j., "Effective Stemming for Arabic information Retrieval," *in proceedings of the Challenge of Arabic for NLP/MT Conference*, Royaume-Uni, 2006.

[17] Khoja S., Garside R., and Knowles G., "A Tag Set For The Morphosyntactic Tagging Of Arabic," *in Proceedings of the Corpus Linguistics Conference*, vol. 13, Special Issue, pp. 341-354, 2001.

[18] Kim J. and Taylor J., "Fast String Matching Using An N- Gram Algorithm," *Journal Of Software: Practice And Experience*, vol. 24, no. 1, pp. 79-88, 1994.

[19] Larkey L., Ballesteros L., and Connell M., "Improving Stemming For Arabic Information Retrieval: Light Stemming And Co-Occurrence Analysis," *in Proceedings of The 25th Annual International ACM SIGIR Conference On Research and Development in Information Retrieval*, Finland, pp. 275-282, 2002.

[20] Larkey L., Ballesteros L., and Connell M., "Light Stemming For Arabic Information Retrieval," *Arabic Computational Morphology*, Springer, 2007.

[21] Mustafa S., "Combining N-Grams And Stemming For Arabic Word-Based Inexact Matching And Term Conflation," *Journal of Information and Knowledge Management*, vol. 4, no. 1, pp. 29-36, 2005.

[22] Nehar A., Ziadi D., and Cherroun H., "Rational Kernels for Arabic Root Extraction And Text Classification," *Journal Of King Saud University-Computer And Information Sciences*, vol. 28, no. 2, pp. 157-169, 2016.

[23] Porter M., "An Algorithm for Suffix Stripping," *Program Journal*, vol.14, no. 3, pp. 130-137, 1980.

[24] Soudi A., Neumann G., and Van-Den-Bosch A., *Arabic Computational Morphology*, Springer, 2007.

[25] Taghva K., Elkhoury R., and Coombs J., "Arabic Stemming Without A Root Dictionary," *in Proceedings of International Conference on Information Technology: Coding and Computing*, Las Vegas, pp. 152-157, 2005.

Yaser Al-Lahham has received the B.S degree from University of Jordan in 1985, the M.S. degree from Arab Academy (Jordan) in 2004, and the PhD in Computer science from Bradford University (UK) in 2009. He is working as an assistant professor in the Department of Computer Science at Zarqa University in Jordan. His research interest includes P2P information retrieval systems, text clustering, and Databases.

Mohammad Hassan has received his BS degree from Yarmouk University in Jordan in 1987, the MS degree from University of Jordan, in 1996, and the PhD degree in computer information systems from Bradford University, UK in 2003. He is working as an associate professor in the department of computer science at Zarqa University in Jordan. His research interest includes information retrieval systems and database systems.

Khawla Al Matarneh has received her BS degree in computer science from Mua'ta University in Jordan in 2004, the MS degree in computer science from Zarqa University, in 2017. Her research interest includes information retrieval systems and database systems.