Arabic Studies' Progress in Information Retrieval

Essam Hanandeh
Computer Information System, Zarqa University,
Zarqa, Jordan

Hayel Khafajah
Computer Information System, Zarqa University,
Zarqa, Jordan

***Abstract*—The field of information retrieval has witnessed tangible progress over the past decades in response to the expanded usage of the internet and the dire need of users to search for massive amounts of digital information. Given the steady increase of Arabic e-content, excellent information retrieval systems must be devised to suit the nature and requirements of the Arabic language. This paper sheds light on the current progress in the field of Arabic information retrieval, identifies the challenges that hinder the progress of this science, and proposes suggestions for further research. This paper uses the descriptive analytical method to examine the reality of Arabic studies in the field of information retrieval and to study the problems that are being faced in this area. Specifically, the previous literature on information retrieval is reviewed by searching the related databases and websites.**

Keywords*—Information retrieval; Arabic information retrieval; Indexing; Query reformulation

I. Introduction

The amount of global digital content has been increased by the continuous information flow from websites, company and government records, e-books, e-newspapers, e-magazines, and other online media. Retrieval systems have become imperative for users to extract information from huge amounts of text, images, and digital sounds. Information retrieval refers to the study of searching for information inside documents or for documents themselves [2]. Such a discipline becomes more important as the number of global Internet users increases and their independence on search engines as a major source for information is strengthened [1].

Text information retrieval involves the processing of natural languages and the retrieval of documents that contain the information that is needed by the user from huge databases. Any classical information retrieval system comprises three basic stages, namely, indexing, query reformulation, and matching. First, all extracted documents are indexed by using the best words or expressions that represent or have an actual indication of each document. Second, the query that is entered by the user to access the required information is reformulated to comply with the information retrieval model as well as to add other keywords or modify the weights of the existent words to achieve better search accuracy. Third, the entered query is matched with the existing index, and the most similar documents are retrieved and arranged in a descending order [2, 5, 22, 15]. By determining how documents are represented in the index, information retrieval models can control how the reception is represented. Many information retrieval models exist, of which the most common models include the Boolean model, fuzzy model, and vector space model [2, 5, 22, 15].

Given the increasing amount of Arabic digital content on the Internet and other electronic devices, the need to create information retrieval systems and engines that pay special attention to the peculiarities of Arabic—the language of the Noble Qur'an and Prophet Mohammad's traditions as well as one of the most widespread Semitic languages in terms of native speakers—continues to increase [7]. Arabic differs from English and other languages in several aspects. First, Arabic text is read and written from right to left. Second, Arabic forms vary based on their position and adjacent letters. Third, the diacritics in Arabic change the pronunciation of letters, meaning, and case of words [9]. Fourth, Arabic is a derivative—rather than inflectional—language with one of the most sophisticated morphological systems. This language divides the stems based on a specific set of weights to develop words of different meanings from the same stem. All these considerations present challenges to the mechanization of the morphological, syntactic, and semantic analyses of the Arabic language and to the retrieval of Arabic texts.

II. Types of Indexing

A. Automatic Indexing

In automatic indexing, an index is built to describe the content of each document in the database in a way that best accelerates and facilitates the search process [2]. This index is any kind of data structure that is used for storing words, keywords, or the general description of any document. An information retrieval system depends on matching the query of the user with all the inputs in the index in order to access the documents that are most similar to the query. The difficulty of indexing documents depends on the processed language. In other words, those languages with sophisticated syntactic and morphological systems, such as Arabic, require highly complicated logarithms [13].

The automatic indexing of Arabic texts enjoys the lion's share of the papers in the field of Arabic text retrieval. This type of indexing is divided into pre-indexing processing, stem-finding-based indexing, stem-making-based indexing, indexing based on stem-making and language rules, dictionary-based indexing, taksir plural indexing, and weighing indexing words.

- Stem-Making-Based Indexing

In stem-making-based indexing, the prefixes and suffixes are extracted from words and the stems are used to index documents. The new words tend to have the same meaning because these affixes are often used to indicate definition, number, sex, coordination, or preposition, which removal will not affect the meaning. Previous studies [15, 18, 19, 20] show that stem-making-based indexing outperforms original-word-

based indexing, stem-finding-based indexing, and stem-making-context-related-based indexing in terms of precession and recall levels. Such high levels are attributed to the very derivative nature of Arabic, which makes the language highly sensitive to stem making [18].

- Indexing Based on Stem-Making and Language Rules

The indexing based on stem-making and language rules is similar to the stem-making-based method, but employs linguistic rules to obtain better results for the stem-making process. A recent study [13,21] shows that this method outperforms the others in terms of stem-making accuracy. Nevertheless, no experiment has been conducted to merge this type of indexing with an Arabic text retrieval system to measure its efficiency.

- Dictionary-Based Indexing

In dictionary-based indexing, each word in the document is indexed by using synonyms [13]. A study on the retrieval of Qur'anic verses shows that this method has a higher retrieval accuracy than the stem-finding-based technique. Another study [18] shows that this method increases the competence of an information retrieval system for Arabic texts by 18%.

- Pre-Indexing Processing (Normalization)

Pre-indexing process is an important stage to obtain the optimal results for the indexing process; this stage involves the removal of diacritics, letters, and stop words that do not have independent meanings [8] as well as the unification of the forms of letters. For instance, the varieties of the Arabic (ألف) letter (أ، ا، إ، آ) are all made [1]. The same applies to the (هاء) varieties (ة، ه), which are both made (ه), and the (ياء) varieties (ي، ئ، ى), which are made (ي) as in [16, 19]. Such shifts are proven successful in improving the retrieval of Arabic texts, which can be attributed to the fact that original texts do not consider the differences between these letters because of the weak Arabic writing language of those who enter such texts.

- Roots-Finding-Based Indexing

In roots-finding-based indexing, the roots are extracted from the document to be used as terms. All words with the same roots will be indexed under the same word even though they may not necessarily have the same meaning. This method has been investigated in many papers [19, 3] and its excellence over original-word-based indexing has also been proven. This technique has achieved high levels of precession and recall in those sets that contain limited or unchangeable numbers of documents, such as those of Qur'anic verses or Prophet Mohammad's traditions. Such high levels are attributed to the fact that this method retrieves all documents that contain any morphological form of the query words, thereby increasing the possibility of finding the required information. However, this technique is impractical in cases of huge and continuously renewed sets, such as those of the Internet. This technique also expands the search scope without providing the user with his/her target.

- Taksir Plural Indexing

Returning the taksir plurals to their original singulars presents a challenge to the Arabic language in general and to the retrieval of Arabic texts in particular. Unlike regular male and female plurals, taksir plurals are not immediately recognized from the text. Various infixes can also be used. Previous research [18] has attempted to address this problem by employing the n-gram technique, but this technique has been proven insufficient. Another study [25] has used a dictionary that lists the singular forms of the taksir plurals to recognize the words. Previous studies have proven that indexing techniques that bring back the taksir plurals to their original singulars outperforms the other indexing techniques.

- Weighing Indexing Words

In weighing indexing words, each term is given a weight that best fits the extent to which the word represents its origin document. Previous research [25,27] has investigated the effects of removing letters or stop words and using various types of weighing indexing words on the retrieval of Arabic texts. The OKAPI BM25 technique and the removal of the stop words can lead to better retrieval results than can the other weighing techniques, such as term frequency-inverse document frequency (tf-idf) and the relevance value of a document with respect to a query that measured by the Kullback-Leibler (KL) divergence between the query model and document model. In addition, when the text is not edited or when no words are removed, the prominent tf-idf method is considered the optimal technique.

Another study [21] explores 12 weighing techniques based on three factors, namely, the number of times the word is repeated in the document, the number of times the stems of such words are found, and the distribution of the word in the document. This technique has been proven efficient in terms of precession and recall.

B. Automatic Query Reformulation

Query reformulation is an information retrieval technique that is applied for adding and/or re-weighing query words to obtain the largest number of matching documents. Query reformulation can be conducted in three ways, namely, relevance feedback, automatic local analysis (inductive query by example), and automatic global analysis [2]. The automatic reformulation of Arabic queries has been investigated in many studies over the past decade.

- Relevance *Feedback Query Reformulation*

In relevance feedback, the user is requested to determine whether the retrieved documents are relevant to his/her query. Accordingly, the query is reformulated by adding words that are mentioned in relevant documents, by removing words that are found in irrelevant documents, or by re-weighing the terms. The new query is entered in the information retrieval system to retrieve another set of documents that may be more relevant. This method is sometimes repeated until the user is satisfied with the results.

In a related study, the user is asked to classify the retrieved documents as relevant or irrelevant. The user is also requested to choose synonyms to the appropriate terms from a dictionary and then include these synonyms in his/her new query. If the added synonyms are highly relevant to the original terms, such an interactive method for investigating the meanings of words

and expanding the query can lead to satisfactory results in terms of precession and recall. However, such results cannot be obtained if the synonyms have a general nature.

Furthermore, an experiment-based study [18] shows that expanding the query by such an interactive way (relevance feedback by the user) outperforms the automatic method (automatic local analysis) in terms of retrieval efficiency. Using either of these methods is better than any using other techniques for reformulating and expanding the query.

- *Automatic Local Context Analysis Query Reformulation*

Automatic local context analysis query reformulation, also called inductive query by example, provides the user an information retrieval system with a set of documents that are either relevant or irrelevant to his/her query. The system then deduces the main words from the relevant documents and sometimes excludes irrelevant words from a query in order to access other relevant documents [14]. However, this method is only employed with frequent queries instead of single-time queries [9].

Authenticity [11] is a major Arabic text retrieval system that is based on the Prophet's traditions. This system identifies the roots of the words that are used in the query and matches them with a roots-finding-based index to produce an initial list of documents. Afterward, automatic local context analysis is used to reformulate the query. After application to one of the queries, the method has yielded 0.66 and 0.80 precession and recall scores, respectively. The success depends on the set of documents to which the method is applied. This method is more appropriate for a highly limited and unchangeable set because the search results can somehow be limited. By contrast, this method is less efficients for larger sets. Specifically, the precession and recall levels are lowered as the scope of the search is significantly expanded.

- Automatic *Global Analysis Query Reformulation*

Unlike the previous two methods, automatic global analysis query reformulation establishes a relation among all terms for all documents in the set and not only between the relevant and irrelevant documents. Most of the techniques attempt to build a dictionary of similarity to determine the relation between terms according to the concept that they represent and not only their simultaneous existence in the same document [2].

Many studies have investigated the application of this method to Arabic text retrieval. For example, the Arab search engine Barq [17] depends on the automatic or manual addition of new query words on three concept dictionaries and on the unification of forms of letters as mentioned above in automatic indexing. This method has increased the precession measure to 75%. Mustafa et al. [28] propose a method for expanding the query by finding synonyms to terms and their derivations. The Neuro-Fuzzy logic has been adopted to obtain the closest derivations to the meaning of the original terms, thereby providing the user with options to expand the query. Researchers have conducted further experiments to prove the efficiency of the method in text retrieval. Another study [7] attempts to expand the query to retrieve information from an Arabic text with or without diacritics. The same method has been applied to the Noble Qur'an by using four types of indexes, namely, index for words with diacritics, index for words without diacritics, root-finding-based index, and synonym-set-based index. He then compares the stem-finding-based index with the query-expansion-based index and finds that the latter outperforms the former in terms of average accuracy.

Another study [28] proposes a modification to the concept-based query expansion—introduced in [30]— to remove the irregular values that are generated by the presence of a very similar word that outshines the less similar ones. This method has improved the retrieval system efficiency by 3.3%.

C. Matching Function Adjustment

In matching function adjustment, the entered query is matched with the index to retrieve documents that are identical to the query. Such documents are called relevant documents that arranged in a descending manner according to their relation to the subject. When designing the matching function, which matches the query with the index, the following must be considered: (1) how to decide whether the provided document is relevant, and (2) how to arrange the relevant documents according to their relevance or ranking [6]. The matching function efficiency depends on several external factors, such as the size of the document set, subject of the document, and culture of the user that has formulated the query [6]. Therefore, unless used in all the information retrieval systems, a particular matching function cannot be proven as successful.

Only few studies have investigated the matching of Arabic texts with the measures of similarity to be used in the field of Arabic information retrieval. One of these studies [28] have explored the efficiency of the n-gram technique in matching and retrieving Arabic text. They have successfully applied such technique with other languages, such as English, because of the highly derivative nature of Arabic, which words also contain infixes. In another study, the n-gram technique is modified to suit the Arabic language. Specifically, the non-consecutive letters of a word are selected and matched them with the letters of other words. In addition to taking the prefixes and suffixes from the stem, the modified technique yields better results than the classical technique. The same technique has been modified by other scholars [8,23] to fit the Arabic language searching in specific locations of the target word. Such modifications aim is to increase the possibility of finding a significant degree of similarity between two words that do not hold the same concept. The modifications outperform the classical methods in terms of precession and recall. These modifications also help find high degrees of similarity among different derivations of the word.

In a recent study [24], researchers build an information retrieval system in Arabic according to the Fuzzy model, believing that this system suits the nature of the Arabic language and can discover the similarity between various synonyms and different sentence structures. This system is based on two dictionaries, namely, one with a matrix that indicates the relation among all words (correlation) and one for synonyms. To determine the similarity between two sentences, the correlation is calculated between each word and each sentence in which the word is found. Afterward, the similarity between the two sentences is calculated. This system

outperforms those information retrieval systems that are based on the Boolean model in terms of precession and recall, thereby proving that the former can detect similarities between similar documents yet requires costly and complicated calculations.

D. Automatic Documents Classification

In the field of information retrieval, if the documents of the same set react similarly to a query [1], then they are classified accordingly. In other words, if one document in a certain set is relevant to a certain query, the rest of the documents in that same set tend to be classified as relevant. Based on the sets of documents to be classified and the aspects of information retrieval to be improved, several applications for automatic classification can be divided into two types. In the first type, the search results are classified in a particular point or in the entire set of documents. In the second type, the classification is performed to improve the interface or experience of the user as well as the efficiency of the search system [1].

Only few studies have classified Arabic documents for the purpose of information retrieval. One of these studies [25] perform a classification based on the Naive Bayes logarithm to create an index of subjects that can facilitate the search process. The documents are divided into five main subjects, namely, sport, business, culture and arts, science, and health. Before the classification process, the diacritics are removed and the stems are identified. The classification accuracy reaches 68.78%. Other scholars [19] propose a logarithm for the automatic classification of Arabic documents by finding those words that cover the main concept of each document subject. Each word is weighed based on the number of times it is repeated and to its locations in the documents. The above classification logarithm enhances the efficiency of the information retrieval system.

In [13] and [28], the efficiency of two logarithms in splitting the text is measured, and these logarithms have been proven successful in both English and Arabic. TextTilling and C99 have excellent application in Arabic, with the former outperforming the latter.

E. Web Page Automatic Search

Crawlers are programs that track hyperlinks on the web, gather pages, and make these pages available to search engines for indexing. These programs are often given URLs or keywords, track the hyperlinks on these webpages, and then move to other pages [6, 25]. Searching in webpages represents a significant challenge because of their large number, which increases on a momentary basis. In addition, given that their contents continue to change, the webpages that are visited earlier must be found and stored to be re-visited and indexed later. The changes in a webpage are unstable and vary according to the type of websites. Webpages can be stored in the following ways [13]:

- *Uniform Policy: All previously indexed webpages are updated whether their contents have been changed.*
- *Proportional Policy: The webpages are updated according to their average change.*
- *Optimal Policy: Only those webpages with trackable changes can be updated.*
- *Curve Fitting Policy: The calculation covers the changes between two consecutive images of the webpage and the number of changes as reflected in the change date.*

The Arabic context remains in its early stage. According to [12,13], Arabic webpages only account for 0.1% of the total webpages, which explains the lack of research on the Arabic language. Another study [5] modifies the curve fitting policy to suit the Arabic language by omitting pronouns, relative pronouns, and prepositions from the content without changing the meaning. They also take the various derivations of the same word with the same meaning. Such modification has reduced time and space, which are important factors in searching for webpages. In another study [13], to search for webpages in Arabic and other languages, a program is distributed to more than one server to enhance speed and efficiency. The speed can reach 160 webpages per second.

III. Conclusion

Information retrieval in Arabic has witnessed tangible progress over the last decade. Specifically, the Arabic document set has provided researchers with a huge number of data. This research has used two sets, first set is published by Saad [29] contains queries and documents that was collected from CNN Arabic website, and second set is BBC Arabic corpus, which has been collected from BBC Arabic website. However, these documents set has several flaws, such as limited syntactic structures, forms of nouns, and verbs as well as many misspelled names of people and non-Arabic places.

Furthermore, given the importance of stem-making for Arabic information retrieval systems, researchers must build an efficient, accurate tool for the stem-making of Arabic words that pay special attention to taksir plurals. Those texts with diacritics must also be reconsidered, and the presence of diacritics must be utilized in disambiguating the meanings of words before starting the indexing process. Differentiation must also be performed between limited, near-constant texts, such as the Noble Qur'an and Prophet Mohammad's Hadith traditions, and huge, continually changing texts, such as webpages. The future work of this research coud be provided and investaged syntactic structures, and forms of nouns of Arabic languge by utilizing disambiguating of words meanings before starting the indexing process.

Acknowledgment

This research is funded by the Deanship of Research and Graduate Studies in Zarqa University /Jordan

Reference

[1] A. Abdelali, J. Cowie, H. Soliman, "Arabic information retrieval perspectives", In Proceedings of JEP-TALN 2004 Arabic Language Processing, 2004.

[2] A.Alhroob, H. Khafajeh , N. Innab, 2013. Evaluation of different query expansion techniques for Arabic text retrieval system. Am. J. Applied Sci., 10: 1018-1024.

[3] M. AL-Kabi, H.Wahsheh, I.Alsmadi, (2014). A Topical Classification of Hadith Arabic Text, IMAN 2014: 2nd International Conference on

Islamic Applications in Computer Science and Technologies, 12th – 13th October 2014, Amman, Jordan, pp. 1-8.

[4] D. Kraft, F. Petry, B. Buckles, T. Sadasivan, "The use of genetic programming to build queries for information retrieval," Evolutionary Computation, 1994 search. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on, 1994, pp. 468-473 vol.1.

[5] D. Ezzat, M. Abdeen, M.F. Tolba, "A Memory Efficient Approach for Crawling Language Specific Web: The Arabic Web as a Case Study," icime, pp.584-587, 2009 International Conference on Information Management and Engineering, 2009.

[6] D. Manning, P. Raghavan, and, H. Schütze, An Introduction to Information Retrieval, Cambridge University Press, 2009.

[7] E. Hanandeh, "SIMILAR THESAURUS BASED ON ARABIC DOCUMENT: AN OVERVIEW AND COMPARISON," International Journal of Computer Science, Engineering and Applications (IJCSEA), Vol.3, No.2, April 2013

[8] E. Hanandeh, K. Mabreh. "EFFECTIVE INFORMATION RETRIEVAL METHOD BASED ON MATCHING ADAPTIVE GENETIC ALGORITHM "Journal of Theoretical and Applied Information Technology, 30 th November 2015 –Vol. 81. No. 3 - 2015

[9] F. Ahmed , A. Nürnberger, "N-grams Conflation Approach for Arabic", ACM SIGIR Conference, 2007.

[10] F. Ataa Allah, S. Boulaknadel, A. El qadi, D. Aboutajdine, "Arabic Information Retrieval System Based on Noun Phrases," Information and Communication Technologies, 2006. ICTTA '06. 2nd, 2006, pp. 1720-1725.

[11] F. Harrag, A. Hamdi-Cherif, E. El-Qawasmeh, "Vector space model for Arabic information retrieval — application to "Hadith" indexing," Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the, 2008, pp. 107-112B.

[12] G. Kanaan, R. Al-Shalabi, M. Sawalha, "Improving Arabic Information Retrieval Systems Using Part of Speech Tagging", Information Technology Journal, vol.4, 2005, pp.32-37.

[13] G. Kanaan, R. Al-Shalabi, M. Ababneh, A. Al-Nobani, "Building an effective rule-based light stemmer for Arabic language to improve search effectiveness," 2008 International Conference on Innovations in Information Technology, Al Ain, United Arab Emirates: 2008, pp. 312-316.

[14] H. Khafajeh, A. Abu-Errub, A. Odeh, N. Youse, (2012) NOVEL AUTOMATIC QUERY BUILDING ALGORITHM USING SIMILARITY THESSAURUS, American Journal of Applied Sciences 9 (9): 1373-1377, ISSN 1546-923.

[15] H. Khafajeh, N. Yousef, (2013) Evaluation of Different Query Expansion Techniques by using Different Similarity Measures in Arabic Documents, International Journal of Computer Science Issues, Vol 10, Issue 4, No 1, July 201, (p.p. 160-166) .

[16] I. El Emary, J. Atwan, "Designing and building an automatic information retrieval system for handling the Arabic data", American Journal of Applied Sciences, 2005.

[17] J. Mayfield, P. McNamee, C. Costello, C. Piatko, A. Banerjee, "JHU/APL at TREC 2001: Experiments in Filtering and in Arabic, Video and Web retrieval", InTREC 2001 Proceedings, 2001.

[18] J. Xu, A. Fraser, R. Weischedel, "Empirical studies in strategies for Arabic retrieval," Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Tampere, Finland: ACM, 2002, pp. 269-274.

[19] L. Larkey, L. Ballesteros, M. Connell, "Light Stemming for Arabic Information Retrieval," Arabic Computational Morphology, 2007, pp. 221-243.

[20] M. Aljlayl, O. Frieder, "On arabic search: improving the retrieval effectiveness via a light stemming approach", CIKM 2002, pp.340-347.

[21] N. Mansour, R.A. Haraty, W. Daher, M. Houri, "An Auto-indexing Method for Arabic Text," Information Processing and Management: an International Journal., vol. 44, 2008, pp. 1538-1545.

[22] N. Yousef, A. Abu-Errub, A. Odeh, H. Khafajeh, AN IMPROVED ARABIC WORD'S ROOTS EXTRACTION METHOD USING N-GRAM TECHNIQUE, Journal of Computer Science 10 (4): 716-719, 2014, ISSN: 1549-3636, © 2014 Science Publications, doi:10.3844/jcssp.2014.716.719 Published Online 10 (4) 2014

[23] P. Pathak, M. Gordon, Weiguo Fan, "Effective information retrieval using genetic algorithms based matching functions adaptation," System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on, 2000, p. 8 pp. vol.1.

[24] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, Wokingham, UK, 1999.

[25] S. Alzahrani, N. Salim, "On the Use of Fuzzy Information Retrieval for Gauging Similarity of Arabic Documents", Proceedings of the 5th Postgraduate Annual Research Seminar, UTM, pp.256-260, 2009.

[26] S. Boulaknadel, B. Daille, A. Driss, "Multi-word term indexing for Arabic document retrieval", Proceedings of the 13th IEEE Symposium on Computers and Communications (ISCC 2008), Marrakech, Morocco 2008.

[27] V. Shkapenyuk, T.Suel, (2002), 'Design and implementation of a high-performance distributed web crawler', In Proc. of the Int. Conf. on Data Engineering.

[28] S. Mustafa, Q. Al-Radaideh, "Using N-Grams for Arabic Text Searching", Journal Of The American Society For Information Science And Technology, vol.55, pp.1002–1007, 2004.

[29] www.sourceforge.net/projects/ar-text-mining

[30] Y. Qiu, H. Frei, "Concept based query expansion". In proceedings of the 16th International ACM SIGIR Conference on R & D in Information Retrieval, ACM Press, New York, 1993, pp. 160-169.