Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Development of Arabic evaluations in information retrieval

Shakir Khan *, Mohammed Alshara

College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

ARTICLE INFO

Article history:
Received 15 January 2019
Received in revised form
5 October 2019
Accepted 8 October 2019

Keywords:
Information retrieval
Arabic information retrieval
Indexing
Query reformulation

ABSTRACT

The field of information retrieval has observed noticeable growth over the past decades in reaction to the prolonged practice of the internet and the dreadful requirement of users to hunt for huge amounts of digital information. Assuming the stable intensification of Arabic e-content, brilliant information retrieval systems must be planned to uniform the nature and needs of the Arabic language. This paper shelters graceful on the present development in the field of Arabic information retrieval finds the trials that delay the development of this learning and proposes recommendations for additional research. This paper practices the imaginative analytical technique to scrutinize the genuineness of Arabic educations in the field of information retrieval and to learn the difficulties that are being confronted in this area. Especially, the earlier literature on information retrieval is reviewed by searching the connected databases and websites.

1. Introduction

The quantity of worldwide digital content has been amplified by the nonstop information run from websites, corporation and government proceedings, e-books, e-magazines, e-newspapers, and further online medium. Retrieval systems have developed into very important for consumers to extort information from vast quantities of text, imagery, and digital sounds. Information retrieval refers to the learning of penetrating for information in papers or for papers themselves (Alhroob et al., 2013). Such a regulation turns out to be more significant as the number of worldwide Internet users enhances and their sovereignty on seek out engines as a key source for information is made stronger (Abdelali et al., 2004).

Text information retrieval contains the handing out of natural languages and the recovery of documents that include the information required by the user from vast databases. Any standard information retrieval system includes three necessary phases, explicitly, indexing, query reformulation, and matching. Foremost, all extorted documents are indexed by utilizing the most excellent words or terminology that stand for or have an authentic suggestion of every document. Subsequent, the query that is penetrated by the user to get access the necessary information is reformulated to act in accordance with the information retrieval representation and to add other keywords or amend the weights of the existing words to attain better search precision. Third, the penetrated query is harmonized with the available index, and the majority related documents are retrieved and organized in a descending order (Alhroob et al., 2013; Ezzat et al., 2009; Yousef et al., 2014; Khafajeh and Yousef, 2013). By formative how documents are characterized in the index, information retrieval representations can manage how the response is represented. Several information retrieval models available, of which the mainly universal models consist of the Boolean model, fuzzy model, and vector space model (Alhroob et al., 2013; Ezzat et al., 2009; Yousef et al., 2014; Khafajeh and Yousef, 2013).

Given the growing quantity of Arabic digital content on the Internet and new electronic devices, the requirement to generate information retrieval structures and engines that pay extraordinary concentration to the customs of Arabic—the words of the Noble Qur'an and Prophet Mohammad's ethnicity and one of the mainly extensive Semitic languages in conditions of native speakers are increasing continuously (Hanandeh, 2013). Arabic differs from English and supplementary languages in different aspects. First, Arabic manuscript is interpreted and written from right to left [RTL]. Second, Arabic appearances differ based on their place and adjoining letters. Third, the diacritics in

* Corresponding Author.
Email Address: shakirkhan2006@gmail.com (S. Khan)
https://doi.org/10.21833/ijaas.2019.12.011
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-7925-9191

Arabic transform the articulation of letters, sense, and container of words (Ahmed and Nijrnberger, 2007). Fourth, Arabic is a plagiaristic—rather than inflectional—language with one of the most difficult morphological systems. This language segregates the branches based on a definite set of weights to build up words of dissimilar meanings from the same branch. All these thoughts present disputes to the computerization of the morphological, syntactic, and semantic investigations of the Arabic language and to the retrieval of Arabic transcripts.

2. Types of indexing

2.1. Automatic indexing

In automatic indexing, a catalog is fabricated to explain the content of every document in the database such that most excellent speeds up and makes possible the search method (Alhroob et al., 2013). This index is any kind of data structure that is utilized to accumulate words, keywords, or the broad explanation of any manuscript. An information retrieval system depends on identical the query of the customer with the entire key in the index in order to get access the documents that are mainly alike to the query. The complexity of indexing documents depends on the procedure language. In supplementary words, those languages with complicated syntactic and morphological systems, for example Arabic, need extremely complex logarithms (Kanaan et al., 2008).

The automatic indexing of Arabic wording takes pleasure in the lions allocate of the documents in the field of Arabic content retrieval. This style of indexing is separated into pre-indexing dealing out, stem-finding supported indexing, stem-making support indexing, indexing supported on stem-making and language rules, dictionary-supported indexing, taksir several indexing, and evaluated indexing terminology.

2.1.1. Stem-building-supported indexing

In stem-building-supported indexing, the prefixes and suffixes are pulled out from terminology and the stop words are utilized to index documents. The novel terms have a tendency to have the identical significance because these fixes are frequently used to point out definition, quantity, sex, synchronization, or preposition, which elimination will not influence the denotation. Earlier studies (Khafajeh and Yousef, 2013; Xu et al., 2002; Larkey et al., 2007; Aljlayl and Frieder, 2002) illustrate that stem-building-support indexing better than original-word-support indexing, stem-discovery-support indexing, and stem-building-context-related-supported indexing in conditions of precession and recall levels. Such high levels are accredited to the extremely copied nature of Arabic, which builds the language extremely responsive to stem building (Xu et al., 2002).

2.1.2. Indexing supported on stem-building and language rules

The indexing supported on stem-building and language rules is parallel to the stem-building-supported technique, but utilizes linguistic rules to attain improved outcomes for the stem-building method. A new study (Kanaan et al., 2008; Mansour et al., 2008) demonstrates that this technique outperforms the others in conditions of stem-building correctness. However, no experimentation has been carried out to combine this kind of indexing with an Arabic wording retrieval system to determine its effectiveness.

2.1.3. Dictionary-supported indexing

In dictionary-supported indexing, all word in the document is indexed with utilizing synonyms (Kanaan et al., 2008). Learning on the retrieval of Qur'anic rhyme demonstrates that this process has privileged retrieval correctness than the stem-discovery-supported method. One more study (Xu et al., 2002) demonstrates that this technique boosts the capability of an information retrieval system for Arabic manuscripts by 18%.

2.1.4. Pre-indexing processing (Normalization)

Pre-indexing development is a significant phase to attain the most favorable results for the indexing method; this phase engages the elimination of diacritics, letters, and stop words that do not include self-determining meanings (Hanandeh and Maabreh, 2015) and the alliance of the appearances of letters. For example, the varieties of the Arabic ('alif) letter in four forms which are all ended (Abdelali et al., 2004). The similar applies to the (haa') diversities are in two forms, which are together made (haa'), and the (yaa') diversities in three forms, which are completed (yaa') as in El Emary and Atwan (2005) and Larkey et al. (2007). With additional same kind of study applies to the (waaw) varieties in two form, which are both made (waaw), (El Emary and Atwan, 2005; Larkey et al., 2007), Such modifications are established fruitful in improving the retrieval of Arabic texts, which can be accredited to the information that innovative texts do not reflect the differences between these letters because of the weak Arabic lettering language of those who writes these kind of texts.

2.1.5. Roots-discovery-supported indexing finding

In roots-discovery-based indexing, the roots are pulled out from the document to be utilized as conditions. All terms with the similar roots will be indexed in the identical word even though they may not essentially have the identical sense. This technique has been examined in several papers (Larkey et al., 2007; Al-Kabi et al., 2015) and its

brilliance over original-text-based indexing has also been confirmed. This method has attained high levels of precession and recollect in those sets that control limited or consistent facts of documents, like those of Qur'anic rhymes or Prophet Mohammad's society. Such sky-scraping levels are recognized to the actuality that this technique retrieves all documents that control several morphological outlines of the query words, thus growing the option of decision the essential information. However, this method is unusable in cases of vast and constantly improved sets, such as individuals of the Internet. These methods also spread out the search range with no providing the consumer with his/her objective.

2.1.6. Taksir plural indexing

Recurring the taksir plurals to their novel singulars shows a confront to the Arabic language in all-purpose and to the retrieval of Arabic wordings particularly. Different usual male and female plurals, taksir plurals are not instantaneously known from the wording. A variety of infixes can also be utilized. Earlier study (Xu et al., 2002) has challenged to address this trouble by utilizing the n-gram method, but this method has been demonstrated inadequate. Another research (Alzahrani and Salim, 2009) has applied a dictionary that records the singular outlines of the taksir plurals to identify the terms. Earlier studies have verified that indexing procedures that carry back the taksir plurals to their novel singulars outperforms the further indexing methods.

2.1.7. Weighing indexing terminology

In weighing indexing terminology, each word is specified a weight that most excellent fits the scope to which the statement corresponds to its source document. Earlier investigation (Alzahrani and Salim, 2009; Shkapenyuk and Suel, 2002) has examined the special effects of eliminating letters or stop words and utilizing different types of weighing indexing terminology on the retrieval of Arabic wording. The OKAPI BM25 method and the elimination of the stop words can guide to enhanced retrieval outcome than can the other weighing methods, for example term frequency-inverse document frequency (tf-idf) and the significance worth of a document concerning a query that considered by the Kullback-Leibler (KL) discrepancy between the query model and document model. Additionally, when the wording is not abbreviated or when no words are detached, the well-known tf-idf technique is measured the most advantageous technique. Another research (Mansour et al., 2008) investigate 12 weighing methods derived from three aspects, specifically, the quantity of times the statement is recurring in the document, the quantity of times the stems of such terminology are found, and the division of the statement in the document. This method has been demonstrated proficient in provisions of precession and recall.

2.2. Automatic query reformulation

Query reformulation is an information retrieval procedure that is useful for totaling and/or re-weighing query vocabulary to attain the major number of identical documents. Query reformulation can be carried out in three customs, that is to say, relevance opinion, automatic local examination (inductive query by model), and automatic worldwide examination (Alhroob et al., 2013). The automatic reformulation of Arabic queries has been examined in various researches over the past decade.

2.2.1. Relevance opinion query reformulation

In relevance opinion, the customer is asked for determining whether the retrieved documents are appropriate to his/her query. Consequently, the query is reformulated by totaling terms that are pointed out in relevant documents, by eliminating words that are found in unrelated documents, or by re-weighing the conditions. The novel query is penetrated in the information retrieval system to recover a new set of documents that may be further relevant. This technique is from time to time recurring until the user is contented with the outcome. In a connected research, the user is inquired to categorize the retrieved documents as related or unrelated. The user is also asked to select synonyms to the suitable terms from a dictionary and then consist of these synonyms in his/her novel query. If the additional synonyms are extremely appropriate to the innovative vocabulary, such an interactive technique for examining the significance of words and growing the query can guide to acceptable outcome in conditions of precession and recall. Conversely, such outcome cannot be achieved if the synonyms have a common nature. Moreover, an experiment-based research (Xu et al., 2002) depicts that getting higher the query by such an interactive technique (relevance feedback by the customer) outperforms the repeated technique (automatic local investigation) in provisions of retrieval effectiveness.

Utilizing either of these techniques is better than any other methods for reformulating and growing the query.

2.2.2. Automatic local framework investigation query reformulation

Automatic local framework investigation query reformulation, also named inductive query by instance, presents the customer an information retrieval system with a position of documents that are either related or unrelated to his/her query. The system then figures out the foremost words from the appropriate documents and occasionally eliminates

unrelated terminology from a query in order to access further related documents (Khafajeh et al., 2012). Though, this technique is only engaged with recurrent queries rather than single-time queries (Ahmed and Nijrnberger, 2007).

Legitimacy (Harrag et al., 2008) is a main Arabic text retrieval system that is found on the Prophet's ethnicity. This system classifies the extraction of the terms that are utilized in the query and equals them with a roots-discovery-supported index to create a preliminary record of documents. Subsequently, automatic local framework investigation is utilized to reformulate the query. Following the application to one of the queries, the technique has yielded 0.66 and 0.80 precession and recall scores, correspondingly. The achievement depends on the set of credentials to which the technique is functional. This technique is more suitable for an extremely incomplete and consistent set as the search outcome can someway be restricted. By distinction, this technique is less competent for superior sets. Particularly, the precession and recall height are lowered as the range of the search is considerably long-drawn-out.

2.2.3. Automatic worldwide investigation query reformulation

Unlike the earlier two techniques, automatic worldwide investigation query reformulation initiated a relation among all conditions for all credentials in the set and not only between the related and unrelated documents. The majority of the procedures challenge to construct a dictionary of resemblance to find out the relation among terms according to the idea that they characterize and not only their immediate survival in the similar document (Alhroob et al., 2013).

Various researches have examined the request of this technique to Arabic wording retrieval. For instance, the Arab search engine Barq (Mayfield et al., 2002) depends on the automatic or manual totaling of novel query terminology on three idea dictionaries and on the association of shapes of letters as pointed out above in repeated indexing. This technique has improved the precession determine to 75%. Mustafa and Al-Radaideh (2004) put forward a technique for growing the query by outcome synonyms to provisions and their beginning. The Neuro-Fuzzy logic has been accepted to acquire the adjoining derivations to the significance of the original vocabulary, thus provided that the customer with choices to enlarge the query. Researchers have performed more experiments to confirm the effectiveness of the technique in text retrieval. One more research (Hanandeh, 2013) challenges to enlarge the query to retrieve information from an Arabic content with or without diacritics. The identical technique has been applied to the Noble Qur'an by utilizing four kinds of indexes, explicitly, index for words with diacritics, index for terminology without diacritics, root-discovery-supported index, and synonym-set-based index. He then weighs against the stem-discovery-supported index with the query-expansion-supported index and locates that the concluding outperforms the earlier in provisions of average accurateness.

Another research (Mustafa and Al-Radaideh, 2004) suggests an adjustment to the concept-supported query expansion—set up in (Qiu and Frei, 1993)—to eliminate the uneven values that are produced by the occurrence of a much related word that outperforms the less related ones. This technique has enhanced the retrieval system effectiveness by 3.3%.

2.3. Matching function modification

In matching function modification, the entered query is coordinated with the index to recover documents that are the same to the query. Such documents are named related documents that prearranged in a downward approach according to their relation to the topic. When scheming the matching function, which equals the query with the index, the subsequent must be measured: (1) how to make a decision whether the supplied document is related, and (2) how to put together the related documents according to their significance or ranking (Christopher et al., 2009). The matching function effectiveness depends on a number of outer aspects, like the mass of the document set, theme of the document, and background of the user that has planned the query (Christopher et al., 2009). As a result, unless utilized in all the information retrieval systems, an exacting matching function cannot be confirmed as winning.

Only the minority researches have examined the similar of Arabic manuscript with the procedures of resemblance to be utilized in the area of Arabic information retrieval. One of these researches (Mustafa and Al-Radaideh, 2004) has discovered the effectiveness of the n-gram method in matching and retrieving Arabic content. They have effectively applied such method with other languages, like English, because of the extremely copied nature of Arabic, which terminology also control infixes. In another research, the n-gram technique is customized to outfit the Arabic language. Particularly, the non-consecutive writing of a statement is chosen and matched them with the letters of other terminology. Additionally, attractive the prefixes and suffixes from the stem, the customized method yields better outcome than the classical method. The equivalent method has been customized by other scholars (Hanandeh and Maabreh, 2015; Pathak et al., 2000) to robust the Arabic language searching in precise locations of the objective word. Such amendments aim is to boost the option of discovery an important degree of resemblance between two words that do not embrace the identical thought. The amendments outperform the traditional techniques in provisions of precession and recall. These amendments also facilitate to come across high degrees of resemblance among dissimilar derivations of the statement.

In a current research (Baeza-Yates and Ribeiro-Neto, 1999), researchers develop an information retrieval system in Arabic with respect to the Fuzzy model, considering that this system costumes the nature of the Arabic language and can find out the resemblance between different synonyms and dissimilar sentence structures. This system is pedestal on two dictionaries, that is, one with matrix that points out the relation amongst all words (correlation) and one for synonyms. To find out the resemblance between two sentences, the association is planned between each word and each sentence in which the statement is found. Subsequently, the resemblance among the two sentences is designed. This system surpasses those information retrieval systems that are footed on the Boolean model in terms of precession and recall, thus proving that the earlier can identify resemblance between similar documents yet needs costly and difficult calculations.

2.4. Automatic documents categorization

In the area of information retrieval, if the documents of the identical set respond alike to a query (Abdelali et al., 2004), then they are categorized consequently. In supplementary words, if one document in a definite set is related to a definite query, the rest of the documents in the same set are subject to be categorized as related. Based on the set of documents to be categorized and the features of information retrieval to be enhanced, numerous applications for automatic categorization can be separated into two kinds. In the first kind, the search outcomes are classified in an exacting point or in the whole set of documents. In the second kind, the categorization is carried out to get better the interface or knowledge of the user and the effectiveness of the search system (Abdelali et al., 2004).

Only few researches have categorized Arabic documents for the point of information retrieval. One of these researches (Alzahrani and Salim, 2009) carry out a categorization based on the Naive Bayes logarithm to generate an index of topics that can ease 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 categorization correctness reaches 68.78%. Other scholars (Larkey et al., 2007) recommend a logarithm for the automatic categorization of Arabic documents by judgment those words that envelop the main idea of each document theme. All word is evaluated based on the number of times it is recurring and to its locations in the documents. The above categorization logarithm improves the effectiveness of the information retrieval system.

In Kanaan et al. (2008) and Mustafa and Al-Radaideh (2004), the effectiveness of two logarithms in splitting the content is calculated, and these logarithms have been confirmed winning in both English and Arabic. Text-Tilling and C99 have brilliant application in Arabic, with the earlier outperforming the latter.

2.5. Web page automatic search

Crawlers are programs that trail hyperlinks on the web, assemble pages, and create these pages accessible to search engines for indexing. These programs are frequently provided URLs or keywords, follow the hyperlinks on these Web-Pages, and then shift to further pages (Christopher et al., 2009; Alzahrani and Salim, 2009). Searching in web-pages characterize an important confront because of their big number, which enhances on a brief basis. Additionally, provided that their contents persist to amend, the web-pages that are visited previously must be set up and stored to be re-visited and indexed soon after. The amendments in a webpage are uneven and differ according to the kind of websites. Web-pages can be stored in the subsequent ways (Kanaan et al., 2008):

- Standardized Policy: All formerly indexed web-pages are efficient whether their contents have been altered.
- Comparative Policy: The web-pages are efficient according to their standard alteration.
- Optimal Policy: Only those web-pages with observable alteration can be reorganized.
- Curve Appropriate Policy: The computation wraps the alterations between two successive images of the webpage and the number of alterations as reproduced in the alteration date.

The Arabic structure remains in its premature phase. In accordance with (Kanaan et al., 2005; 2008), Arabic web-pages only account for 0.1% of the whole web-pages, which enlightens the lack of study on the Arabic language. Another research (Ezzat et al., 2009) changes the curve fitting policy to costume the Arabic language by leaving out pronouns, comparative pronouns, and prepositions from the content without altering the sense. They also receive the different origins of the same statement with the same sense. Such alteration has concentrated time and space, which is significant, aspects in searching for web-pages. In one more research (Kanaan et al., 2008), to search for web-pages in Arabic and further languages, a program is spread to further than one server to develop speed and effectiveness. The speed can get to 160 web-pages per second.

3. Conclusion

Information retrieval in Arabic has observed touchable development over the previous decade. Particularly, the Arabic document set has given researchers with a giant number of data. This study has utilized two sets, first set is published online by Saad (2014) includes queries and documents that was composed from CNN Arabic website, and second set is BBC Arabic mass, which has been composed

from BBC Arabic website. However, these documents set has numerous faults, such as restricted syntactic structures, outlines of nouns, and verbs and various misspelled names of natives and non-Arabic locations.

Furthermore, specified the significance of stem-making for Arabic information retrieval systems, researchers must construct a well-organized, precise tool for the stem-making of Arabic words that disburse extraordinary concentration to taksir plurals. Those contents with diacritics must also be re-evaluated, and the existence of diacritics must be employed in disambiguating the significance of words before opening the indexing development. Segregation must also be carried out between restricted, near-constant texts, such as the Noble Qur'an and Prophet Mohammad's Hadith society, and vast, frequently altering texts, such as web-pages. The future work of this study could be granted and explored syntactic structures and kinds of nouns of Arabic language by using disambiguating of language meanings before beginning the indexing procedure.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abdelali A, Cowie J, and Soliman HS (2004). Arabic information retrieval perspectives. In the 11th Conference on Natural Language Processing, Journes d'Etude sur la Parole-Traitement Automatique des Langues Naturelles, Fez, Morocco: 391-400.

Ahmed F and Nijrnberger A (2007). N-grams conflation approach for Arabic text. SIGIR'07 iNEWS07 workshop, Amsterdam, Netherlands: 1-8.

Alhroob A, Khafajeh H, and Innab N (2013). Evaluation of different query expansion techniques for Arabic text retrieval system. American Journal of Applied Sciences, 10(9): 1018-1024. https://doi.org/10.3844/ajassp.2013.1018.1024

Aljlayl M and Frieder O (2002). On Arabic search: Improving the retrieval effectiveness via a light stemming approach. In the 11th International Conference on Information and knowledge Management, ACM, McLean, USA: 340-347. https://doi.org/10.1145/584792.584848

Al-Kabi MN, Wahsheh HA, Alsmadi IM, and Al-Akhras AMA (2015). Extended topical classification of hadith Arabic text. International Journal on Islamic Applications in Computer Science and Technology, 3(3): 13-23.

Alzahrani SM and Salim N (2009). On the use of fuzzy information retrieval for gauging similarity of Arabic documents. In the Second International Conference on the Applications of Digital Information and Web Technologies, IEEE, London, UK: 539-544. https://doi.org/10.1109/ICADIWT.2009.5273835

Baeza-Yates R and Ribeiro-Neto B (1999). Modern information retrieval. Addison Wesley, Boston, USA.

Christopher DM, Raghavan P, and Schutze HS (2009). Introduction to information retrieval. Cambridge University Press, Cambridge, UK.

El Emary I and Atwan J (2005). Designing and building an automatic information retrieval system for handling the Arabic data. American Journal of Applied Sciences, 2(11): 1520-1525. https://doi.org/10.3844/ajassp.2005.1520.1525

Ezzat D, Abdeen M, and Tolba MF (2009). A memory efficient approach for crawling language specific web: The Arabic web as a case study. In the 2009 International Conference on Information Management and Engineering, IEEE, Kuala Lumpur, Malaysia: 584-587. https://doi.org/10.1109/ICIME.2009.105

Hanandeh E and Maabreh K (2015). Effective information retrieval method based on matching adaptive genetic algorithm. Journal of Theoretical and Applied Information Technology, 81(3): 446-452.

Hanandeh ES (2013). Similar thesaurus based on Arabic document: An overview and comparison. International Journal of Computer Science, Engineering and Applications, 3: 2. https://doi.org/10.5121/ijcsea.2013.3201

Harrag F, Hamdi-Cherif A, and El-Qawasmeh E (2008). Vector space model for Arabic information retrieval—Application to "Hadith" indexing. In The 2008 First International Conference on the Applications of Digital Information and Web Technologies, IEEE. Ostrava, Czech Republic, 107-112. https://doi.org/10.1109/ICADIWT.2008.4664328

Kanaan G, Al-Shalabi R, Ababneh M, and Al-Nobani A (2008). Building an effective rule-based light stemmer for Arabic language to improve search effectiveness. In The 2008 International Conference on Innovations in Information Technology, IEEE, Al Ain, UAE: 312-316. https://doi.org/10.1109/INNOVATIONS.2008.4781687

Kanaan G, Al-Shalabi R, and Sawalha M (2005). Improving Arabic information retrieval systems using part of speech tagging. Information Technology Journal, 4(1): 32-37. https://doi.org/10.3923/itj.2005.32.37

Khafajeh H and Yousef N (2013). Evaluation of different query expansion techniques by using different similarity measures in Arabic documents. International Journal of Computer Science Issues (IJCSI), 10(4): 160-166.

Khafajeh H, Abu-Errub A, Odeh A, and Yousef N (2012). Novel automatic query building algorithm using similarity thesaurus. American Journal of Applied Sciences, 9(9): 1373-1377. https://doi.org/10.3844/ajassp.2012.1373.1377

Larkey LS, Ballesteros L, and Connell ME (2007). Light stemming for Arabic information retrieval. In The Arabic Computational Morphology, Springer, Dordrecht, Netherlands: 221-243. https://doi.org/10.1007/978-1-4020-6046-5_12

Mansour N, Haraty RA, Daher W, and Houri M (2008). An auto-indexing method for Arabic text. Information Processing and Management, 44(4): 1538-1545. https://doi.org/10.1016/j.ipm.2007.12.007

Mayfield J, McNamee P, Costello C, Piatko C, and Banerjee A (2002). JHU/APL at TREC 2001: Experiments in filtering and in Arabic, video, and web retrieval. In the 10th Text Retrieval Conference (TREC'01), NIST Special, Gaithersburg, USA.

Mustafa SH and Al-Radaideh QA (2004). Using n-grams for Arabic text searching. Journal of the American Society for Information Science and Technology, 55(11): 1002-1007. https://doi.org/10.1002/asi.20051

Pathak P, Gordon M, and Fan W (2000). Effective information retrieval using genetic algorithms based matching functions adaptation. In the 33rd annual Hawaii International Conference on System Sciences, IEEE, Maui, USA. https://doi.org/10.1109/HICSS.2000.926653

Qiu Y and Frei HP (1993). Concept based query expansion. In the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Pittsburgh,

USA: 160-169. https://doi.org/10.1145/160688.160713

Saad M (2014). Arabic computational linguistics. Available online at: https://bit.ly/2KmeeJH

Shkapenyuk V and Suel T (2002). Design and implementation of a high-performance distributed web crawler. In the 18th International Conference on Data Engineering, IEEE, San Jose, USA: 357-368. https://doi.org/10.1109/ICDE.2002.994750

Xu J, Fraser A, Weischedel R, and Weischedel R (2002). Empirical studies in strategies for Arabic retrieval. In the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, Tampere, Finland: 269-274. https://doi.org/10.1145/564422.564424 **PMCid:PMC2732469**

Yousef N, Abu-Errub A, Odeh A, and Khafajeh H (2014). An improved Arabic word's roots extraction method using n-gram technique. Journal of Computer Science, 10(4): 716-719. https://doi.org/10.3844/jcssp.2014.716.719