EAI Endorsed Transactions
on Internet of Things
Review Article EAI.EU

A Review on Recent Arabic Information Retrieval Techniques

Abdelkrim Aarab*, Ahmed Oussous and Mohammed Saddoune

LIM, Hassan II University of Casablanca, Casablanca, Morocco abdelkrim.aarab-etu@etu.univh2c.ma, ahmed.oussous@fstm.ac.ma, mohammed.saddoune@gmail.com

Abstract

Information retrieval is an important field that aims to provide a relevant document to a user information need, expressed through a query. Arabic is a challenging language that gained much attention recently in the information retrieval domain. To overcome the problems related to its complexity, many studies and techniques have been presented, most of them were conducted to solve the stemming problem. This paper presents an overview of the Arabic information retrieval process, including various text processing techniques, ranking approaches, evaluation measures, and some important information retrieval models. The paper finally presents some recent related studies and approaches in different Arabic information retrieval fields.

Keywords: Information Retrieval, Arabic, Natural language processing, Indexing, Ranking, Evaluation

Received on 31 july 2022, accepted on 29 September 2022, published on 25 October 2022

Copyright © 2022 Abdelkrim Aarab *et al.*, licensed to EAI. This is an open access article distributed under the terms of the CC BY-NC-SA 4.0, which permits copying, redistributing, remixing, transformation, and building upon the material in any medium so long as the original work is properly cited.

doi: 10.4108/eetiot.v8i3.2276

1. Introduction

Searching for information is one of the most daily activities of web users. Nowadays, the web contains a huge amount of data (mostly unstructured) and is continuing to grow exponentially in various forms. This has led to more research in finding effective tools and techniques to get the right and relevant information in an easier and faster way. The field of Information Retrieval (IR) was born in the 1950s out of this necessity; it was initiated by focusing on text and text documents only[1]. Today, information retrieval is related to multiple domains, such as text classification, image retrieval, speech, expert search, spam filtering, question-answering, cross-language IR, and many others.

IR can be defined as: "Information retrieval (IR) which is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information's need from within large collections (usually stored on computers)"[2].

The two main goals of an Information retrieval system are effectiveness and efficiency. Effectiveness refers to the need to find and retrieve the most relevant documents, whereas efficiency refers to the need to find those documents quickly[3].

The present paper is a preliminary attempt to review the recent works related to Arabic Information retrieval trends and techniques. The initial section presents a brief overview of the Arabic language and the challenges of Arabic IR. The second section examines the Arabic text's preprocessing and indexing steps. In the third section, we outline the current approaches in Arabic IR. Finally, some of our conclusions are drawn in the last section.

2. The Arabic Language

Arabic is one of the vital Semitic languages that originate from the Arabian Peninsula that then spread into Western Asia and North Africa as the Muslim World expanded. It is today an official language in 23 countries, which makes it the native language of over 447 million people as well as the liturgical language for over a billion Muslims all over the world. Statistics show that out of the estimated 447,572,891 people in the world that speak Arabic, 53.0% are Internet

* Corresponding author. Email: abdelkrim.aarab-etu@etu.univh2c.ma

users. Moreover, considering how far the Arabic language had gone throughout the years, it is now established as the fourth top language on the web, not to mention how fast it grew in the process.

Figure 1, illustrates the top ten languages used on the web recently. As shown, the number of Arabic-speaking internet users has grown by 9,348% in the last twenty years, and today they constitute 5.2% of all Internet users. The Arabic language can be then categorized into three classifications. Coming first is the Classic Arabic (CA) which is found in pre-Islamic poetry, old literature, and mostly the Holy Quran. Secondly, the Modern Standard Arabic (MSA) which is used by the Arabic governments, newspapers, and modern literature. Thirdly, the Arabic Dialect (AD) which is used especially in social media and has many variations[4].

2.1. Challenges of Arabic IR

Arabic has 28 letters that are written from right to left (RTL), fifteen of which come with dots to distinguish them from each other (for example ب،ت،ث), and they change shape depending on their placements in words[5]. For example, the letter "ب" (b) is written this way "بـ" if it is the first letter in a word, this way "ـبـ" if it is in the middle, and this way "ـب" or "ب" if it is the last letter. Additionally, the Arabic script can be written with diacritics, in other words, vowels which are not part of Arabic alphabets, instead they are written as marks (ـَ ـُ ـِ). Diacritics give more clarification to the meaning and pronunciation of Arabic words. As an example, < ولد > (wld) can have two meanings depending on the presence of diacritics; < وُلد > (wulida) means "born" in English, or < وَلَد >(walad) which means "boy". Thus, the absence of short diacritics can ultimately alter the meaning of a word.

Compared to most European languages including English, Arabic morphology has much more complexity. Excluding scientific terminologies and nomenclature, English has only 400,000 keywords which have a total of 1.3 million words, in the other hand, Arabic has about 5 million words (almost four times more words than English) that are derived from around 11,300 roots. The single root can generate tens to hundreds of derivatives. Most roots are composed of three letters or Tri-literal[6].

Arabic is a rich language in terms of morphology thus it is a heavily inflected language. It has two types of morphology: derivational and inflectional. The latter is used to inflect gender, tense, etc. from a given stem. For example, the stem < قرأ> (read) could have the present tense inflection < يقرأ > (he read), the plural inflection <يقرؤون> (they read), and the feminine inflection < قرأت > (she read). Regarding Derivational morphology, it can lead to changing the meaning of a word entirely. For instance, given the stem <كتب> (wrote) could have the derivations < كاتب > (writer), < كتابة > (writing), and <مكتبة > (library). The added morphemes could be at the beginning (affixes), at the middle (infixes), or at the end (suffixes).[7], [8]

A single Arabic word can relatively take the form of a whole sentence when translated to other languages. For example, "فسيكفيكهم" means "He will suffice you against them" in English. That's why the cursive nature of the Arabic language makes it difficult to recognize the stem from the added morphemes. Such complex features present a big challenge for effective retrieval. Overall, this paper focuses on Arabic IR because it has been less studied compared to English language. Note, the major Arabic IR techniques are briefly explained in the next section.

3. The Information Retrieval Process

The aim of this section is to provide an overview of the information retrieval process, which can be divided in two major phases: offline and online. **Figure 2,** illustrates the processing techniques in the offline phase. The online processing techniques on the other hand, are presented in **Figure 3**.

The search for text is the most common information retrieval application. Different steps of Arabic text preprocessing information retrieval are described in this section.

3.1. Text Acquisition

The text acquisition process aims to recognize and acquire all the documents that are bound to be searched, which are called a collection. Some collections already exist for testing purposes like the Arabic Gigaword Fifth Edition which contains 334,6167 texts of newspaper articles. Although, in the case of a search engine, text acquisition is required to crawl the web or any other sources of information to build a collection or corpus.

3.2. Text Transformation

The text transformation phase is required as a preliminary task in Arabic IR, which aims to identify the optimal form of the term to be indexed to achieve the best retrieval performance. This covers steps like removing punctuation and diacritics, unification of various shapes of some Arabic letters like ('إ', 'أ' and 'آ') which are replaced by 'ا', ('ئ' and 'ؤ') by 'ء', 'ى' by 'ي', and 'ة' becomes 'ه'. Here follows some major preprocessing techniques.

Tokenization

Tokenization is the process of splitting text into single words[9]. One possible way to do so is to split the document sentences into a list of tokens using white spaces[7]. However, there are non-segmented languages, like Chinese, which does not have white spaces between words[10].The generated tokens might be any contiguous sequence of letters or numbers. Thus, a token is an instance of a sequence of characters, and is a candidate for an index entry, after further processing.

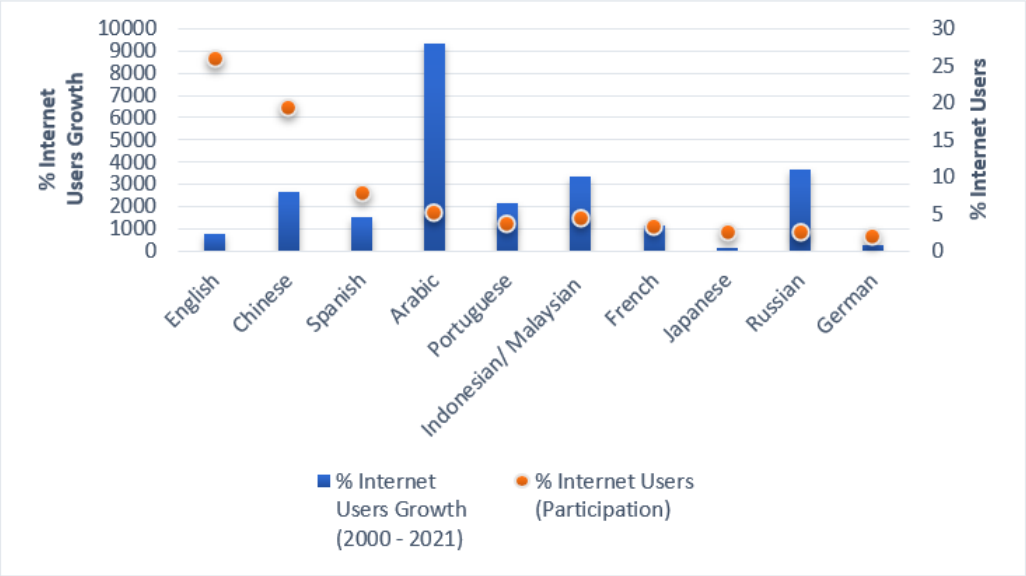

Figure 1. Top ten languages used on the web

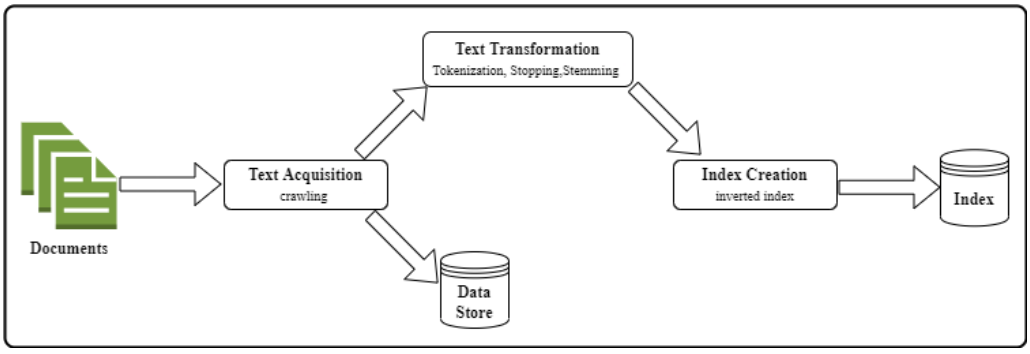

Figure 2. IR Offline process

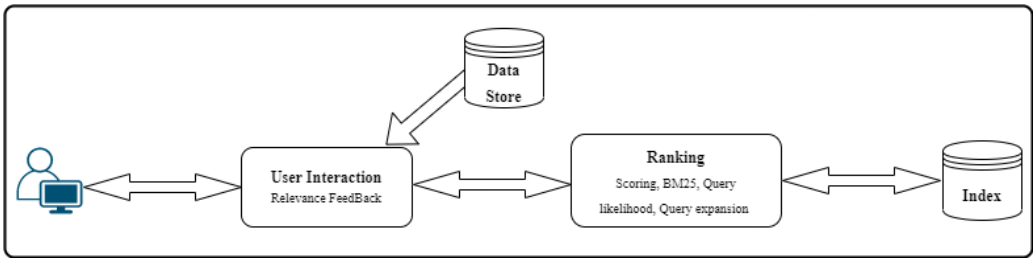

Figure 3. IR Online process

Stopping

Stop-words removal is a necessary stage in Arabic IR. Stop words are the common function words which carry no meaning when used individually[9], it only helps form phrases such as '**على**' <on>, '**و**' <and>, '**في**' <in>, etc. One of the Arabic stop words issues is that they may contain prefixes in addition to suffixes. Removing stop words reduces the storage spaces required to store identified tokens. Not to mention, many studies showed that removing Stop-words improve the efficiency and effectiveness of Arabic IR systems[5], [11].

Stemming

Stemming is an important phase in text processing[8]. It aims to reduce morphological variations of words to a common stem. In Arabic IR, two main stemming approaches are used: stem-based and root-based. The stem-based approach, also known as light stemming, extracts the words stem after separating its affixes using a predefined list containing a variety of prefixes and suffixes[12]. The root-based approach, namely heavy stemming, however, returns the root of derived words from predefined list of samples after the words prefixes and suffixes were removed[12]. In contrast to light stemmers, root-based stemmers produce a different meaning of a word based on the same root. For example, when applying a light stemmer on the word 'مكتباتنا'<our libraries>, it results in the word 'مكتبة'<library> while the root word 'كتب'<write> is returned when a heavy stemmer is used. Thus, many studies show that the use of light stemmer is better than other stemming types[6]. However, when coming to irregular plurals, most of the Arabic light stemmer algorithms fails to get the right stem[13]. Recently, the broken plural rule algorithm (BPR) was designed to override the irregular broken plural in Arabic language processing[8].

Lemmatization

Lemmatization takes a more complex approach in text processing; It aims to regroup semantically related words, and it is proved to be beneficial in the areas of Arabic information retrieval[11], [14]. However, in Arabic, the use of lemmatization is more difficult task due to the morphological complexity of the language itself, and the absence of short vowels in most existing Arabic documents[15].

3.3. Indexing

The index construction is an essential phase[16] that refers to the process of listing and sorting all collection terms with a list of all documents that contain those terms. To represent term-document relationships, an inverted index is used[3]. The latter contains two main elements: dictionary and postings list. The Dictionary is a list of all collection terms sorted alphabetically with their document frequency, while the postings list is a sorted set of documents IDs that contains the term in addition to the term position in the case of positional indexes[2].

3.4. User Interaction

Users contribute greatly to the IR process because, in the end, it is the user information need based on a user query that should be satisfied. Also, user interactions are often used in reranking tasks. Although keywords are used to specify the query topic, some additional techniques are used to satisfy the information need of users such as the reformulation and expansion of queries.

Query transformation

Query transformation refers to the processing techniques that should be applied to the user query to generate index terms that are used to see if they match the document terms or not. The basic treatment is to use some of the same text transformation techniques described above, namely, Tokenization, stopping, and stemming. Some complicated techniques are also used such as query suggestion and spelling checking[3].

3.5. Ranking

Ranking is the process of computing and assigning a score to each document matching a query, with the aim to order those relevant documents because, the top ranked documents are the most likely to satisfy user's query. Here follows some main ranking techniques.

Scoring

Scoring is the process of calculating and storing terms weights in lookup tables. Terms weights refer to the importance of terms in documents as there are many terms weighting functions. One of the best-known weighting schemes is TF-IDF which is used to represent the importance of a term in a document effectively[9]. TF refers to Term Frequency which is the number of occurrences of a term t in document d, while IDF refers to Inverse Document Frequency which captures the importance of infrequent terms in the documents collection because, rare terms are more informative than frequent terms. IDF of a term t is given by:

$$idf_t = log_{10} \frac{N}{df_t} \quad (1)$$

Where N is the number of the collection documents, and df_t refers to document frequency of term t, which is the number of documents that contain term t, thus $df_t \leq N$.

The TF-IDF weighting term can be expressed as:

$$w_{t,d} = tf_{t,d} * log_{10} \frac{N}{df_t} \quad (2)$$

The scoring function of a document-query pair is defined as:

$$score(q,d) = \sum_{t \in q \cap d} w_{t,d} \quad (3)$$

BM25

BM25 Term Weighting is another popular term weighting scheme[12]. BM 25 is an optimized version of TF-IDF, which doesn't consider the length of documents that may be exploited by spammers to cheat search engines by stuffing unnecessary keywords. As an example, a million-word document can contain all the query terms, but may not satisfy the user information need instead, a short document that contains all the query terms can be more relevant. Thus, BM25 is more realistic comparing to TF-IDF's term weighting function as every different term in the document have been represented when the relevance is computed[17].

The BM25 term weighting function can be expressed as:

$$w_{t,d} = \frac{(k_1+1)\,tf_{t,d}}{k_1\left((1-b)+b\frac{L_d}{\bar{L}}\right)+tf_{t,d}} * log\left(\frac{N-df_t+0.5}{df_t+0.5}\right) \quad (4)$$

Where $L_d$is the length of a document d, and $\bar{L}$ is the average documents length in a collection. k_1 and b are parameters and the experiments performed on TREC showed that 1.2 is the ideal value for k_1 , whereas the typical value of b is 2[3].

Query likelihood

The query likelihood model's basic idea is to suppose that the user has in mind some relevant documents and picks some keywords from these documents to use them to formulate the query[17]. To rank the documents, the probability of a document d given the query q is calculated using Bayes' Rule:

$$\mathrm{P(d|q)} = \frac{P(q|d)P(d)}{P(q)} \quad (5)$$

P(d) and P(q) are ignored because they are the same for all the documents. Thus, the P(d|q) is equivalent to P(q|d) which can be estimated as:

$$\mathrm{P}(q|d) = \prod_{i=1}^{n} P(q_i|d) \quad (6)$$

Where n is the length of a query q, and q_i is a query term. The $P(q_i|d)$ can be then estimated as:

$$P(q_i|d) = \frac{tf_{q_i,d}}{|d|} \quad (7)$$

Where $|d|$ is the length of a document d and $tf_{q_i,d}$ is the term frequency of the query term q_i in a document d.

Query expansion

QE is an approach where the original query is extended. In fact, sometimes queries inserted by the users are not a good representation of their information needs, thus, more similar semantical terms are appended to the main query, to retrieve more related documents. Many techniques have been developed to handle user queries, and to improve information retrieval performance. To select the expansion terms many approaches were developed namely:

(i) Thesaurus-based methods: a thesaurus is a data structure that is built manually (e.g., WordNet[ii]) or automatically, and lists words in a group of synonymous. The query is expanded then by adding the most related synonym of each query term.
(ii) Query Logs: are used to extract the best related terms of queries to use them as expansion terms.
(iii) Relevance Feedback (RF): involves the user by showing him a list of initial results and based on the user feedback (the user marks some documents either as relevant or non-relevant), the initial query is improved by adding related terms.
(iv) Pseudo Relevance Feedback (PRF) use the top ranked documents to choose the top terms (based on term weights) as expansion terms automatically without involving the user [18].

Classical query expansion techniques depend on statistical models such as the latter described above, namely TF-IDF, and BM25. Here follows the four types of QE: Manual Query Expansion (MQE), Automatic Query Expansion (AQE), Interactive Query Expansion (IQE), and Hybrid Query Expansion (HQE)[19]–[21].

3.6. Retrieval Models

Since the beginning of Information Retrieval, various retrieval models have been proposed. There are three main IR models approaches; the oldest is the binary approach, which is the simplest one, then there is the vector space approach, and finally the probabilistic approach[7].

Boolean model

The Boolean IR model is the earliest and simplest IR model that uses the set theory and Boolean logic operators, namely AND, OR and NOT to join query terms. Many search systems still use the Boolean IR model, such as Mac OS X Spotlight, email, library catalog, and legal search. Although The Boolean model returns documents that exactly satisfy the Boolean query, however, it does not rank or classify documents. A document is either matching or not matching the query[17].

Vector Space model

VSM is one of the widely used models until the end of the last century[17]. Documents are represented as vectors as well as queries in an n-dimensional space, while "n" refers to the number of terms in a collection. Thus, each dimension corresponds to a term. Documents and queries are highlighted by a real-valued vector of TF-IDF weights. Documents are ranked according to their proximity to the query in this n-dimensional space, or by the cosine similarity, which is the computing of the cosine of the angle between the query and every document in the collection after the normalization of vectors length.[17], [22]

Okapi BM25 model

One of the most used models is BM25 which is based on probabilistic model. BM25 is basically an extension of the binary independence model. Many studies that were conducted on different IR collections confirmed that its use has led to very significant results[12].

3.7. Evaluation

Evaluating IR systems require many criteria, including the ability to retrieve relevant documents, the capacity to avoid irrelevant documents, response times, how retrieval performance is achieved, and techniques to improve performance effectiveness[23]. Effectiveness Evaluation is

[ii] https://wordnet.princeton.edu/

the ability to define the retrieval effectiveness for a given query[17], and it has two main approaches: set-based and rank-based measures.

Set-based

Set-based is destined for evaluating unranked sets, and it has two primary measures, the first one is Precision and the other one is Recall.

Precision is the proportion of the retrieved documents that are in fact relevant to the user query[17]. Therefore, Precision indicates the ability to retrieve documents that are most relevant.

Recall refers to the fraction of all relevant documents that were found and retrieved[17]. Thus, Recall indicates the ability to retrieve all relevant items in a corpus.

F-measure refers to the weighted average of both Precision and Recall[9].

Rank-based

The rank-based approach is more used in IR systems since most IR models generate ranked results. Here follow some Rank-based measures.

Precision @K measure involves computing precision at fixed level of retrieved results. Note, K is a predefined rank position which is mostly 10 or 20[2].

Average Precision (AP) covers the compute of precision results average that corresponds to rank positions where a relevant document was retrieved.

Mean Average Precision (MAP) is the most frequently used measure in research papers, and it refers to the average of the average precisions, which means that the AP of each query are computed and then divided by the number of queries.

4. Current approaches in Arabic IR

Recently Arabic information retrieval gained many attentions. Many studies have been devoted to overcoming the problems related to the complexity of the Arabic language. This section discusses recent advances in Arabic information retrieval field.

The use of text preprocessing techniques is an essential phase, the authors of [11] demonstrated the impact of Stop Words removal, stemming, and lemmatization, on the efficiency of three different text classification algorithms: NB, SVM, and DT J48. The obtained results showed that the combination of the three preprocessing techniques, leads to better improvement of the three text classification algorithms (+15.57% to +28.73%). In a different study, [24] introduced AMIR or Arabic Morphology Information Retrieval, which is a new technique for extracting Arabic stems. The main contribution consists in applying a set of rules to find the right root or stem of the words that will be indexed. The obtained results confirm the ability of AMIR to ameliorate Arabic stemmer and increase retrieval performances. The authors have also reported that the proposed stem algorithm outperforms UCENE, FARASA and non-stem methods. In another work, [25] introduced an approach for the Arabic information retrieval using Bi-gram query expansion. The experimental results showed that the proposed approach outperforms the stem-based method in terms of precision and recall. More recently, [19] demonstrated the effectiveness of the use of a hybrid semantic query expansion approach, which combines statistical and semantic methods, to generate expansion terms closely related to the main query. The results proved that the value of accuracy has increased significantly in terms of information retrieval. To deal with term mismatch, [12] proposed an approach to incorporate word embedding (WE) semantic similarities. The main idea consists in using a predefined list of similar words to extend the scoring function, which let similar words to match the initial query words. The experiments realized with three different neural word embedding models confirm that the developed extensions improve the baseline of selected bag-of-words models. In another work, [26] introduced an approach to integrate word embedding (WE) similarity into different Pseudo-relevance feedback (PRF) models. The obtained results confirmed that the use of their method increased the baseline IR model MAP by 22% and RI by 68%.

The research presented in [22] proposed a framework for Arabic Documents Information retrieval (ADIR) functions to supply an infrastructure in the aim of supporting the creativity of data sets and OCR systems. The obtained results tested with three various OCRing services and four Arabic documents corpora, showed that the proposed solution outperforms all the classification accuracy rate on testing images. In another work, [9] compared six different IR models (HLM, DFRee, DFR_BM25, DLM, Js_KLs, and TF-IDF) in CLIR system. They also compared three different query expansion techniques (Chi-square, KL, and Bo1), and investigated the suitable length of a query for effective retrieval in the CLIR system using the Quranic dataset. The authors concluded that the use of KL query expansion method with a maximum length of 4 terms in a query attains significant results. Regarding retrieval models, the results showed that Js-KLs outperforms all the others. To improve Arabic light stemming, [13] presented Dlight which is a new stem-based approach. The results of the experiments that were conducted on TREC 2002 dataset, show that the proposed Dlight stemmer outperforms the rest of the Arabic stemmers, namely ARLST, Light10 and Condlight. More recently, the authors of [8] designed a new stemming algorithm to override the irregular broken plural in Arabic language processing. The broken plural rule algorithm (BPR) offers many roles to obtain the right roots of the irregular Arabic plural words. The authors have also improved the original ISRI stemmer method by applying the BPR algorithm to it. The experimental results showed the ability of the proposed approach to stem words with high performance, thus helped in improving the performance of a root-based Arabic stemmer.

Table 1, summarizes some recent approaches developed in Arabic information retrieval field.

Table 1. Recent Arabic Information retrieval studies

Reference	Data Sources (type, size, public or private)	Approaches	Feature Extraction and selection	Evaluation Measures	Obtained results
[25]	Arabic Holy Quran corpus, 6236 documents, 77,430 words, 14,662 tokens,	Query expansion	Bi-Gram	precision and recall	MAP 78.35%
[12]	Arabic TREC 2001/2002 data set, 75 topics	Query expansion Word embeddings (WEs)	CBOW	Precision	MAP 36.34 % P10 51.60 %
				Recall	76.46 %
			Skip-gram	Precision	MAP 36.2 % P10 51.47 %
				Recall	75.61 %
			Glove	Precision	MAP 36.41% P10 52.3%
				Recall	76.58 %
[19]	Arabic corpus (BBC, CNN, and Al-Jazeera), 6464 documents, 48,305 words	Hybrid Query Expansion (HQE)	WordNet, term frequency, Word2Vec	Precision	47.28 %
				Recall	53.06 %
[26]	Arabic TREC 2001/2002 collection, 383,872 documents	Query expansion word-embedding-based Pseudo-relevance feedback (PRF)	Glove	Precision	MAP 41.11 P@10 55.07
			Skip-gram	Precision	MAP 41.26 P@10 54.67
			CBOW	Precision	MAP 41.21 P@10 55.07
[24]	EveTAR(2016) dataset, 59,732 documents	Arabic Morphology Information Retrieval (AMIR)	BM25	Precision	MAP 34% P@10 63% P@20 59%
			LM with Dirichlet smoothing	Precision	MAP 32% P@10 60% P@20 56%
[22]	Datasets of 16,800 Arabic letters	OCRing segmentation and recognition			Varies from 83% to 94%
[9]	Quranic dataset (Tanzil)	Cross-Language Information Retrieval (CLIR) Query Expansion	DLM, DFRee DFR_BM25, HLM, Js_KLs, TF-IDF		MAP@5 76.4%
[11]	Arabic text corpus, 300,000 articles	Text preprocessing	Stop Words removal, Stemming, Lemmatization	10-fold cross-validation	NB: 93.47% SVM: 94.91% DT J48: 90.81%
[13]	Arabic dataset TREC2002	light-based stemming		Precision	60%
				Recall	79%
				F-measure	68%
				ICF	81%
[8]	Arabic dataset TREC2002	stemming	ISRI + BPR stemmer	Precision	79%
				Recall	85%
				F-measure	82%
				ICF	90%

5. Conclusion

In recent years many research efforts were developed in the field of Arabic information retrieval, most of them were conducted to solve the stemming problem[27]. Although Arabic language is ranked as the fourth top language on the web and the fastest growing language in the last decades, research on this language is limited, and further research work is still required due to the morphological complexity of the language itself and the lack of Arabic high-quality and well-structured linguistic and semantic resources. In this review we have described some challenges of Arabic IR, presented some Arabic IR techniques and issues. Given the complexity of Arabic morphology, normalization of text documents is required as a preliminary phase task. In future work, we are interested to compare and combine more retrieval techniques to improve Arabic information retrieval effectiveness.

References

[1] D. Harman, "Information retrieval: The early years," Foundations and Trends in Information Retrieval, vol. 13, no. 5. Now Publishers Inc, pp. 425–577, 2019. doi: 10.1561/1500000065.

[2] C. D. Manning, P. Raghavan, and H. Schütze, "An Introduction to Information Retrieval".

[3] W. Bruce Croft Donald Metzler Trevor Strohman, "Search Engines Information Retrieval in Practice."

[4] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, "Arabic natural language processing: An overview," Journal of King Saud University - Computer and Information Sciences, vol. 33, no. 5, pp. 497–507, Jun. 2021, doi: 10.1016/J.JKSUCI.2019.02.006.

[5] K. Darwish and W. Magdy, "Arabic information retrieval," Foundations and Trends in Information Retrieval, vol. 7, no. 4, pp. 239–342, 2013, doi: 10.1561/1500000031.

[6] IEEE Computer Society., 2011 IEEE GCC Conference and Exhibition : GCC : ... took place February 19-22, 2100 in Dubai, UAE. IEEE Computer Society, 2011.

[7] Z. Alyafeai, M. S. Al-shaibani, M. Ghaleb, and I. Ahmad, "Evaluating Various Tokenizers for Arabic Text Classification," Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.07540

[8] H. Alshalabi, S. Tiun, N. Omar, E. abdulwahab Anaam, and Y. Saif, "BPR algorithm: New broken plural rules for an Arabic stemmer," Egyptian Informatics Journal, Feb. 2022, doi: 10.1016/j.eij.2022.02.006.

[9] A. A. Taan, S. U. R. Khan, A. Raza, A. M. Hanif, and H. Anwar, "Comparative Analysis of Information Retrieval Models on Quran Dataset in Cross-Language Information Retrieval Systems," IEEE Access, vol. 9, pp. 169056–169067, 2021, doi: 10.1109/ACCESS.2021.3126168.

[10] S. Ibrihich, A. Oussous, O. Ibrihich, and M. Esghir, "A Review on recent research in information retrieval," in Procedia Computer Science, 2022, vol. 201, no. C, pp. 777–782. doi: 10.1016/j.procs.2022.03.106.

[11] A. el Kah and I. Zeroual, "The effects of Pre-Processing Techniques on Arabic Text Classification," International Journal of Advanced Trends in Computer Science and Engineering, vol. 10, no. 1, pp. 41–48, Feb. 2021, doi: 10.30534/ijatcse/2021/061012021.

[12] A. el Mahdaouy, S. O. el Alaoui, and E. Gaussier, "Improving Arabic information retrieval using word embedding similarities," International Journal of Speech Technology, vol. 21, no. 1, pp. 121–136, Mar. 2018, doi: 10.1007/s10772-018-9492-y.

[13] H. Alshalabi, S. Tiun, N. Omar, F. N. AL-Aswadi, and K. Ali Alezabi, "Arabic light-based stemmer using new rules," Journal of King Saud University - Computer and Information Sciences, 2021, doi: 10.1016/j.jksuci.2021.08.017.

[14] E. H. Nfaoui, Jāmiʿat Sīdī Muḥammad ibn ʿAbd Allāh. Faculty of Sciences Dhar El Mahraz, IEEE Computer Society, and Institute of Electrical and Electronics Engineers, ISCV'17 : 2017 Intelligent Systems and Computer Vision (ISCV) : April 17-19, 2017, Faculty of Sciences Dhar El Mahraz (FSDM), Fez, Morocco.

[15] A. A. Freihat, M. Abbas, G. Bella, and F. Giunchiglia, "Towards an Optimal Solution to Lemmatization in Arabic," in Procedia Computer Science, 2018, vol. 142, pp. 132–140. doi: 10.1016/j.procs.2018.10.468.

[16] M. A. Abderrahim, M. Dib, M. E. A. Abderrahim, and M. A. Chikh, "Semantic indexing of Arabic texts for information retrieval system," International Journal of Speech Technology, vol. 19, no. 2, pp. 229–236, Jun. 2016, doi: 10.1007/s10772-015-9307-3.

[17] V. N. Gudivada, D. L. Rao, and A. R. Gudivada, "Information Retrieval: Concepts, Models, and Systems," in Handbook of Statistics, vol. 38, Elsevier B.V., 2018, pp. 331–401. doi: 10.1016/bs.host.2018.07.009.

[18] S. Dahir and A. el Qadi, "A query expansion method based on topic modeling and DBpedia features," International Journal of Information Management Data Insights, vol. 1, no. 2, Nov. 2021, doi: 10.1016/j.jjimei.2021.100043.

[19] H. ALMarwi, M. Ghurab, and I. Al-Baltah, "A hybrid semantic query expansion approach for Arabic information retrieval," Journal of Big Data, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00310-z.

[20] Y. H. Farhan, M. Mohd, and S. A. M. Noah, "Survey of Automatic Query Expansion for Arabic Text Retrieval," Journal of Information Science Theory and Practice, vol. 8, no. 4, pp. 67–86, 2020, doi: 10.1633/JISTaP.2020.8.4.6.

[21] M. N. Asim, M. Wasim, M. U. G. Khan, N. Mahmood, and W. Mahmood, "The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval," IEEE Access, vol. 7, pp. 21662–21686, 2019, doi: 10.1109/ACCESS.2019.2897849.

[22] H. M. Al-Barhamtoshy, K. M. Jambi, S. M. Abdou, and M. A. Rashwan, "Arabic Documents Information Retrieval for Printed, Handwritten, and Calligraphy Image," IEEE Access, vol. 9, pp. 51242–51257, 2021, doi: 10.1109/ACCESS.2021.3066477.

[23] A. Omar and M. Aldawsari, "Lexical Ambiguity in Arabic Information Retrieval: The Case of Six Web-Based Search Engines," International Journal of English Linguistics, vol. 10, no. 3, p. 219, Apr. 2020, doi: 10.5539/ijel.v10n3p219.

[24] A. Alnaied, M. Elbendak, and A. Bulbul, "An intelligent use of stemmer and morphology analysis for Arabic information retrieval," Egyptian Informatics Journal, vol. 21, no. 4, pp. 209–217, Dec. 2020, doi: 10.1016/j.eij.2020.02.004.

[25] I. Moawad, W. Alromima, and R. Elgohary, "Bi-Gram Term Collocations-based Query Expansion Approach for Improving Arabic Information Retrieval," Arabian Journal for Science and Engineering, vol. 43, no. 12, pp. 7705–7718, Dec. 2018, doi: 10.1007/s13369-018-3145-y.

[26] A. el Mahdaouy, S. O. el Alaoui, and E. Gaussier, "Word-embedding-based pseudo-relevance feedback for Arabic information retrieval," Journal of Information Science, vol. 45,

no. 4, pp. 429–442, Aug. 2019, doi: 10.1177/0165551518792210.

[27] A. el Mahdaouy, E. Gaussier, and S. O. el Alaoui, “Should one use term proximity or multi-word terms for Arabic information retrieval?,” Computer Speech and Language, vol. 58, pp. 76–97, Nov. 2019, doi: 10.1016/j.csl.2019.04.002.