The Impact of Online Indexing in Improving Arabic Information Retrieval Systems

Tahar Dilekh,
Computer Science Department, University of Batna 2, Batna 05078, Algeria
E-mail: tahar.dilekh@univ-batna2.dz

Saber Benharzallah
Computer Science Department, University of Batna 2, Batna 05078, Algeria
LINFI Laboratory, University of Biskra, Biskra 07000, Algeria
E-mail: s.benharzallah@univ-batna2.dz

Ali Behloul
LaSTIC Laboratory, University of Batna 2, Batna 05078, Algeria
E-mail: a.behloul@univ-batna2.dz

Keywords: online indexing, offline indexing, semi-automatic indexing, Arabic keywords extraction, Arabic information retrieval system

Received: April 17, 2018

This paper suggests a new type of indexing Arabic Language text that contribute to improving the quality of IRS. The proposed method of indexing belongs to semi-automatic category of indexing and consists of two types. The first type conducts an online indexing and the output of this process give a rise to a Partial index. The second type – under this method- is an offline indexing and the output of this process leads to a General index. We illustrate application and the performance of this new method of indexing using an Arabic text editor and Information Retrieval tool developed and designed for this purpose. We also illustrate the process of building a new form of Arabic corpus appropriate to conduct the necessary experiments. Our findings show that the online indexing model successfully identifies the descriptors most relevant to the document. In addition, this model is more efficient as it helps minimizing index storage size, consequently, improving the response time of the different requests. Finally, the paper proposes a solution to issues and deficiencies Arabic language processing suffers from, especially regarding corpora building and information retrieval evaluation systems.

Povzetek: V prispevku je predlagan nov način indeksiranja arabskih besedil z namenom izboljšanja jezikovno-računalniških operacij.

1 Introduction

Recent developments in the internet technology made information abundant, which made it highly available to users. On the other hand, the vast availability of information made it particularly challenging for users to obtain and find relevant and useful information. In this context, Information Retrieval Systems (IRS) have emerged as a tool to address this problem.

IRS consists of two stages: the 'indexing' and the 'search' stages. In the first stage, the descriptors are extracted from documents and prepared to facilitate and accelerate the search process in the second stage. In general, the indexing stage consists of three types. First, manual indexing, in which the descriptors selection process is performed by a human expert. Second, the automatic indexing where the descriptors are automatically extracted from documents, and finally, the semi-automatic indexing (or supervised indexing). This latter provides automated assistance to the expert.

Currently, IRS benefit from the indexing processes, most of which remains under-performing in the extraction of accurate descriptors that contribute to improving the quality of these systems including extracting the semantic of these descriptors. This remains a challenging task of automatic indexing that often requires human intervention to choose the appropriate descriptors. This is because of several reasons including the ambiguity of language, the power of language to transfer thoughts from one mind to another and the dynamic nature of language.

While the literature consists of many studies concerning various natural languages, there are relatively fewer studies on Arabic language, where the complex grammatical and morphological features of this language make the task of automatic processing even more challenging. Thus, this paper suggests a new type of indexing to contribute to improving the quality of IRS. The proposed method of indexing belongs to semi-automatic category of indexing and consists of two types. The first type conducts an online indexing where one document is the indexing unit. This type of indexing refers to the indexing process that begins directly after the writing of each unit ends, which allows to assist human expert (author of text) to select Arabic appropriate descriptors to improve the search results. The output of this process give a rise to a Partial index. The second type – under this method- is an offline indexing, which refers to the process of indexing based on the collection of textual documents available from different corpora. The output of this process leads to a General index.

We also illustrate implementing and the performance of this new method of indexing using an Arabic text editor developed and designed to allow for an online semi-automatic indexing system and Information Retrieval tool that contains an offline automatic indexing system. We also illustrate the process of building a new form of Arabic corpus appropriate to conduct the necessary experiments.

Thus, this study contributes to two key areas of the literature. First, it offers applications of some tools such as SIRAT[1] and OIRDA[2] that have been developed to show the extent to which the integration of online semi-automatic indexer into text editors is effective in improving indexing, and thus improving the precision of IRS. Second, the study is conducted on Arabic texts, which contributes to the enrichment and development of Arabic language processing tools.

The remainder of the paper is organized as follows. Section 2 offers an account of the main developments and recent advances of Arabic documents indexing literature. Section 3 identifies the main characteristics of Arabic language followed by an illustration of the proposed semi-automatic system in Section 4. Section 5 and 6 illustrate implemented applications and analyze the results of the conducted experiments respectively. Section 7 concludes.

2 Literature review

We being with a review the main literature of Arabic documents indexing, and identify the challenges facing this research area. We categorize the literature according to the most commonly used approach. We then present some work related to the automatic Arabic keyword extraction, which helps to improve the quality of Arabic indexing systems.

2.1 Arabic documents indexing

Various studies have proposed different methods for Arabic documents indexing. However, to the best of our knowledge, all of these studies focused on manual and automatic indexing. This prevented us from comparing the existing methods to that proposed in this paper. This paper proposes various automatic indexing techniques according to the following approaches: linguistic, statistical, semantic, and hybrid.

2.1.1 The linguistic approach

The linguistic approaches consist of a morphological and syntactic analysis of the document based on the grammatical rules and relationships between the different textual units. The methods of this approach are widely used in Arabic natural language processing due to the reliability of syntactic and semantic recognition algorithms. Saadi et al. [1] proposed knowledge extraction systems, based on a deep linguistic analysis and using a domain ontology to extract the semantic content, they have achieved promising results, but reveal other problems in need of careful investigation.

Mansour et al.[2] proposed a method mainly based on morphological analysis and on a technique for assigning weights to words. The morphological analysis uses a number of grammatical rules to extract candidate index words. The weight assignment technique computes weights for these words relative to the container document. The weights are based on how spread are the words in a document and not only on their rate of occurrence. The experimental results carried out for a number of texts have demonstrated the advantage of their auto-indexing method.

Al Molijy et al. [3] proposed and implemented a method to create and index for books written in Arabic language using the syntactic analysis. The process depends largely on text summarization and abstraction processes to collect main topics and statements in the book automatically.

This approach offers good results in specific situations, such as determining the exact meaning of a vague word as expressed in the sentence; the name is gold, but the verb is gone, but remains less able to match other approaches, given the complexity of the Arabic language.

2.1.2 The statistical approach

Statistical approaches are mostly based on statistical techniques. A variety of these approaches have been developed to extract descriptors (terms) and study their occurrence in a document, or even in the corpus.

The frequency distribution of words has been a key object of study in statistical approach for the past decades. This distribution approximately follows a simple mathematical form known as Zipf's law. According to this law, words occur according to a systematic frequency distribution such that there are few very high-frequency words that account for most of the text and many low-frequency words. We very briefly mention some of the places where this law affects research in our study:

- Zipf's Law tells us how much text we have to look at and how precise our statistics have to be to achieve what level of expected error [4].
- Zipf's Law also provides a base-line model for expected occurrence of target terms and the answers

[1] The Arabic text editor SIRAT (Semantic Information Retrieval of Arabic Texts) is an application that we have developed to conduct experiments on semantic Arabic information retrieval domain.

[2] It is an indexing and retrieval program for Arabic texts, we have developed in Java. OIRDA is abbreviation of the French sentence (Outil d'Indexation et de Recherche dans les Documents Arabes) i.e. Indexing and retrieval tool for Arabic documents.

to certain questions may provide considerable information about its role in the corpus [5]: what does it mean to ask if a word is significant in a corpus, beyond mere occurrence or relative probability? What is the range of the semantic influence of a word in a corpus? What does the pattern of occurrences contribute to our assessment of its relevance in the corpus? [6]

- Zipf's Law provides a basis for evaluating parsers and taggers [7]. Again we summarize the potential role in the form of a series of questions: How does a language model developed on one corpus transfer to another? How do we translate performance estimates on a few test corpora to estimates for the language as a whole? How do differences in register, genre and medium affect the utility of a system, and how do we compensate for these differences? [6]

The Term Frequency–Inverse Document Frequency (TF-IDF) method is also one of the statistical approaches that provides a good representation of the weight of corpora words whose document size is homogeneous. Several alternatives have been proposed for the TF-IDF method, which has become the subject of many comparative studies.

The feasibility of this approach also depends on the process of extracting the root/stem of each word, according to root-based approach; or stem-based approach; in order to overcome the polymorphism of the word.

Several studies have shown that the process of stemming of the word from its prefixes and suffixes is more useful for Arabic information retrieval systems than in other approaches.

Researchers adopted various statistical methods and techniques in the indexing process [8] [9] [10] [11] [12] [13] [14] [15] [16].

In conclusion, these methods, considered as simple to implement, are efficient and perfectly tolerant of large masses of documentary. On the other hand, the hypothesis considering the words as independent units generates a loss of semantic information. The resulting indexes may generate polysemy problems and deviate from the general context of the document [17].

2.1.3 The semantic approach

This approach aims, on the one hand, to reduce the ambiguity of the words meaning and, on the other hand, allows to extract the semantic relations between these words. Thus, texts are represented focuses on the unit of meaning rather than simple words. Semantic relationships can also be calculated using methods that evaluate the amount of information between words.

Researchers [18] have integrated semantic process into an Internet search engine and used several techniques (Harman, Croft, and Okapi) to evaluate the performance of this engine. In a recent study [19] [20] have exploited the lexical base of Arabic WordNet in an IRS in order to index the collection of documents and query of the user. Others[21], introduced a query expansion approach using an ontology built from Wikipedia pages in addition to other thesaurus to improve search accuracy for Arabic language.

This approach provides the best semantic cover for the documents due to relies on semantic resources (dictionaries, anthologies or others). However, it remains restricted by the type of resource used and its ability to describe the words of the text being processed.

2.1.4 The hybrid approach

Several researchers [22] [23] [24] [25] have experimented with different combinations of linguistic, statistical, and semantic methods, taking the advantages of each method in an attempt to overcome their shortcomings and to improve the process of indexing by extracting hidden information in a document. These approaches often led to better results than those obtained through the use of standard methods.

Despite the positive results of this approach, it suffers from the problem of complexity, depending on the integration of other approaches.

2.2 Extraction Arabic keywords

Keywords (descriptors) are a subset of words or phrases that can describe the meaning of a document, where several natural language processing applications can benefit from keywords. Unfortunately, most documents do not contain these words. On the other hand, adding high-quality keywords manually is costly, time-consuming, and error-prone. Therefore, this domain has emerged to develop novel algorithms and systems designed to extract keywords automatically.

[26] presented the KP-Miner (Keyphrases-Miner) system to extract keyphrases from both English and Arabic documents of varied length. This system does not need to be trained on a particular document set in order to achieve its task (i.e. unsupervised learning). It also has the advantage of being configurable as the rules and heuristics adopted by the system are related to the general nature of documents and keyphrases. In general, Experiments and comparison studies with widely used systems suggest that KP-Miner is effective and efficient.

[27] introduced AKEA, a keyphrase extraction -unsupervised- algorithm for single Arabic documents. They relied on heuristics that collaborate linguistic patterns based on Part-Of-Speech (POS) tags, statistical knowledge and the internal structural pattern of terms. They employed the usage of Arabic Wikipedia to improve the ranking of candidate keyphrases by adding a confidence score if the candidate exists as an indexed Wikipedia concept. Experimental results have shown that the performance of AKEA outperforms other unsupervised algorithms as it has reported higher precision values.

[28] presented a keyword extraction system for Arabic documents using term co-occurrence statistical information. In case the co-occurrence of a term is in the biasness degree, then the term is important and it is likely to be a keyword. The biasness degree of the terms and the set of frequent terms are measured using $\chi 2$. Therefore, terms with high $\chi 2$ values are likely to be keywords. This

technique showed an acceptable performance compared to other techniques.

[29] presented a supervised learning technique for extracting keyphrases of Arabic documents. The extractor is supplied with linguistic knowledge to enhance its efficiency instead of relying only on statistical information such as term frequency and distance. An annotated Arabic corpus is used to extract the required lexical features of the document words. The knowledge also includes syntactic rules based on part of speech tags and allowed word sequences to extract the candidate keyphrases. The experiments carried out show the effectiveness of this method to extract Arabic keyphrases.

[30] presented a framework for extracting keyphrases from Arabic news documents. It relies on supervised learning, Naïve Bayes in particular, to extract keyphrases. The final set of keyphrases is chosen from the set of phrases that have high probabilities of being keyphrases.

Various experiments have shown the effectiveness of these methods to extract Arabic keywords in varying percentages. However, while supervised techniques are costly and limited by the type of language resources used, unsupervised techniques suffer from the best semantic cover for the documents.

3 Characteristics of the Arabic language

The complex grammatical and morphological features of the Arabic language make the task of automatically processing more difficult. Among these features, we highlight the following:

- Arabic scripts have diacritics to represent the short vowels, which are marks above or below the letters. However, these diacritics have been disappearing in most contemporary writings, and readers are expected to fill in the missing diacritics through their knowledge of the language. The absence of diacritics from contemporary Arabic texts makes the automatic processing a difficult task.
- Morphological analysis is a complex procedure because Arabic is an agglutinative language. For example, the word "أفاستسقيناكموها" (*did we ask you- plural- for water to her (it)*) is one of the longest words in the Arabic language dictionaries. It consists of 15 letters and 9 diacritics. Its root is the verb "سقى" (*to water*). We add to the word the prefix "است" to become "استسقى" (*he asked for water*). Adding a subject pronoun, the word becomes "استسقينا" (*we asked for water*). Then we add the indirect object pronoun to become "استسقيناكم" (*we asked you – plural- for water*), and we add the direct object to become "استسقيناكموها" (*we asked you – plural- for water for her (it)*) Next, we add "F" of appeal (ف الاستئناف) and "A" of question (أ الاستفهام) to become a fully-meaningful phrase: "أفاستسقيناكموها؟" (*did we ask you- plural- for water to her (it)?*).
- Arabic is a highly inflectional and derivational language where many of the nouns and verbs are derived from the same root. This latter is based on more than 150 patterns, which makes them more complex and difficult to handle.

4 Semi-automatic indexing system

As emphasised in the introduction above, we have designed and developed a semi-automatic indexing system that is based on:

1. An Online semi-automatic indexing of Arabic documents (Figure 1).
2. An Offline automatic indexing of Arabic corpus (Figure 5).

4.1 Online semi-automatic indexing system

This system consists of three units: a unit for automatic indexing, a unit for the automatic extraction of keywords and a unit for updating partial index of a document after the intervention of the human expert to select the relevant keywords.

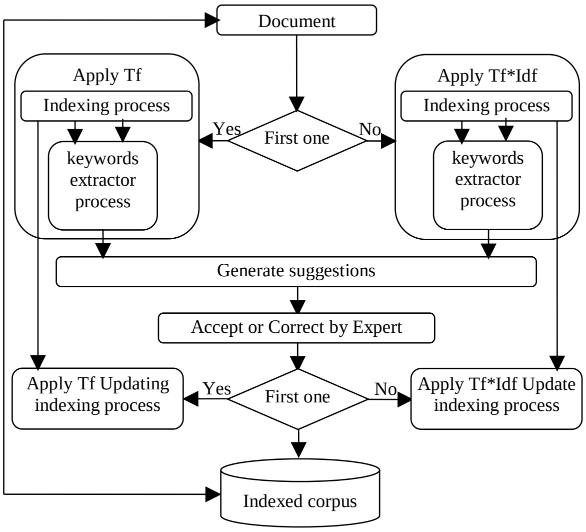

Document
Apply Tf
Indexing process
keywords extractor process
Yes
First one
No
Apply Tf*Idf
Indexing process
keywords extractor process
Generate suggestions
Accept or Correct by Expert
Apply Tf Updating indexing process
Yes
First one
No
Apply Tf*Idf Update indexing process
Indexed corpus

Figure 1: Online semi-automatic indexing system of Arabic documents.

In addition, we integrated our online indexing system to an Arabic text editor (Figure 6) that we developed for the purpose of testing and running experiments. We also created an Arabic corpus in a new format (Figure 7) that allows us running the necessary experiments.

4.1.1 Automatic indexing unit

Indexing is the process of representing the given text into the list of informative terms, which reflects its content in order to optimize speed and performance in finding relevant documents for a search query.

The automatic indexing of Arabic texts had dominated most of the research literature in Arabic text retrieval. In our study, we followed the approach due to [25] to create the index with some modifications, which we discuss in the next section. This method has proved to be effective in improving the process of indexing Arabic documents.

4.1.1.1 Encoding

The corpus and queries can be encoded differently, making them incomparable. In order to standardize the documents with the queries, we must reuse converting tools between different encodings systems. Thus, everything would be converted into UTF-16 encoding in our case, because it allows the representation of letters and symbols in a wide range of languages, including Arabic.

4.1.1.2 Normalization

Normalization involves the following steps:

- Remove punctuation;
- Remove diacritics (primarily weak vowels);
- Remove the Tatweel '-'.
- Replace the 'إ' or the 'أ' initial by Alif nu 'ا';
- Replace the 'آ' by the 'ا';
- Replace the 'ىء' of order by the 'ئ';
- Replace the 'ى' final by the 'ي';
- Replace the 'ة' final by the 'ه'.

4.1.1.3 Removing stop words

The removal of stop words has the advantage of reducing the number of indexing terms and may reduce the recall rate (i.e. the proportion of relevant documents returned by the system to all relevant documents). We use a list of stop words to remove stop words.

4.1.1.4 Stemming

We used a hybrid method, as proposed by [25], to extract the roots of the words and use them as index terms. This combines the application of three previously used techniques, which deal with three key issues related to Arabic stemming including affix removal proposed by [31], dictionaries [32] and morphological analysis[33]. This method has been found to be effective in indexing process compared to other methods.

4.1.1.5 Term frequency and weighting

Several statistical measure are available to assign weights to words of a document in a corpus. Currently, TF-IDF is one of the most popular term-weighting procedure. TF-IDF value increases proportionally to the number of times a word appears in the document and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general.

In our study, we used TF-IDF that combine the definitions of term frequency and inverse document frequency, to produce a composite weight for each term in each document. The TF-IDF weighting procedure assigns a weight to term t in document d given by

$$tf - idf_{t,d} = tf_{i,j} * idf_i$$

where

- $tf_{i,j}$: the number of times that term i occurs in document j.
- $idf_i = \log \frac{|D|}{|\{d_i : t_j \in d_j\}|}$
- $|D|$: total number of documents in the corpus.
- $|\{d_i : t_j \in d_j\}|$: number of documents where the term t appears (i.e., tf (t, d) $\neq 0$).

Our automatic indexing unit deals differently with the first document added to the corpus (Figure 2). Since there are no documents available prior to the first document to compute $tf - idf_{t,d}$, we only count a $tf_{i,j}$ value.

The automatic indexing unit constructs a *partial index* for every document of every corpus. The output of this unit is a *partial index* for each document (Figure 2). The main motivation behind constructing *partial indexes* is to allow the expert intervention in the creation of index later.

```
Indexing function pseudo code
Input:
        Document $d_i$ ϵ corpus
Output:
        $Index_i$ // partial index
Algorithm
For each token in $d_i$ loop
        Encoding ();
        Normalize ();
        Removing_stop_words ();
        Stemming ();
        If (tf_type = tf) then
                Weighting(tf)
        Else
                Weighting(tf-idf);
        End
        Stored tf for the term = token
End loop.
Add $d_i$ to $Index_i$.
```

Figure 2: Automatic indexing algorithm.

4.1.2 Automatic keyword extraction unit

We have adopted a simple method of extracting keywords as long as the human expert is responsible for the final decision making regarding the acceptation or modification of the appropriate keywords for the document being processed (see the example in figure 3).

Instructions	Execute?	If no, why?
1. ***Input:*** في الأمم المتحدة أن ... (In the united nations that …)	-	
2. ***Selected word from the result of the indexing module*** ... في الامم المتحدة أن ...	Yes	
3. ***Add 1st right word*** ... في الامم المتحدة أن ...	No	Stop word
4. ***Add 1st left word*** ... في الامم المتحدة أن ...	Yes	
5. ***Add 2nd left word*** ... في الامم المتحدة أن ...	No	Stop word
6. ***Output:*** الأمم المتحدة (The united nations)		

Figure 3: Automatic keyword extraction example.

The automatic keyword extraction unit (Figure 4) proposes the list of candidate words. This list is limited to twelve keywords, each consisting of at most five words. These words are extracted in two stages:

In the first stage, we adopt the results of the automatic indexing unit, where we retrieve the index words with the highest weights. Then, we add, if possible, to each index word, from original text, two nearest neighbor words on the right and two others on the left while ensuring that this five-word string does not contain Arabic punctuation marks in between words. Otherwise, we just take the number of words between two punctuations. We also give priority to a noun phrase or nominal phrase by setting terms for the candidate words in the following order:

- Words that begin with "ال" letters and end with " ,"ة ي"" or "ء" letters.
- Words that begin with "ال" letters.
- Words that end with "ي" ,"ة" or "ء" letters.
- Ordinary words.

In the second stage, we propose to the human expert twelve key words arranged in descending order, after which the human expert would accept or modify the suggestions generated by the automatic keyword extraction unit.

```
Keywords_Extract function pseudo code
Input:
            Document $d_i$ ϵ corpus
Output :
            Keywords [ ]
Algorithm
For j = 1 to 12 loop
       word ← Paratial_Index.canditat_word[j];
       word ← From_Original_text (word);
       if (Setting_terms (fst_right_word))
              word ←  word + fst_right_word;
       if (Setting_terms (snd_right_word))
              word ←  word + snd_right_word;
       if (Setting_terms (fst_leftt_word))
              word ←  fst_left_word + word;
       if (Setting_terms (snd_lest_word))
              word ←  snd_left_word + word;
     Keywords [i] ← word
End loop.
```

Figure 4: Automatic keyword extraction algorithm.

4.1.3 Unit of updating partial index

The role of this unit is to update a partial index of a document. The expert's opinions are accounted for by updating the weights of the selected index words and assigning to them higher values. This phase concludes with the integration of this partial index into the document, and saving it to an object file in order to exploit it later.

4.2 Offline indexing system of building and updating general index

The role of this system is to build and update a general index based on partial indexes of several corpora (Figure 5). It retrieves all documents indexes (partial indexes) that created by the online semi-automatic indexing system, and merges them into a single general index. It also updates this index whenever necessary.

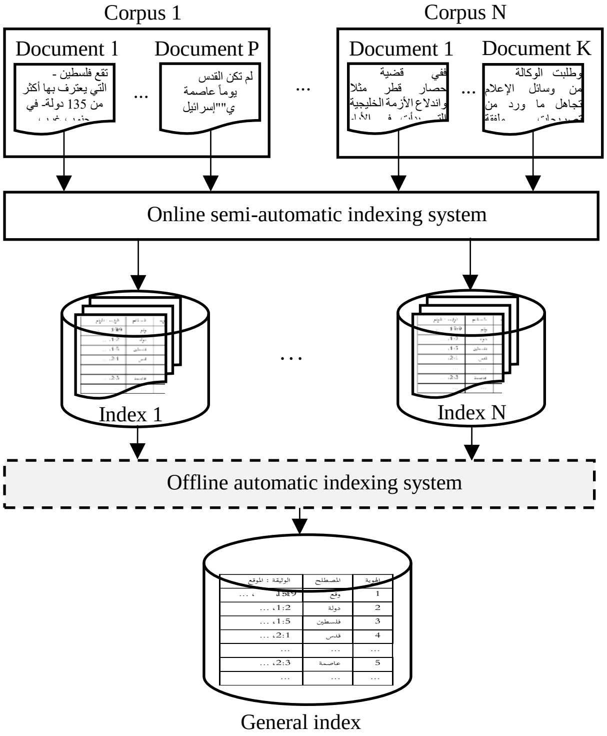


Figure 5: Offline automatic indexing system.

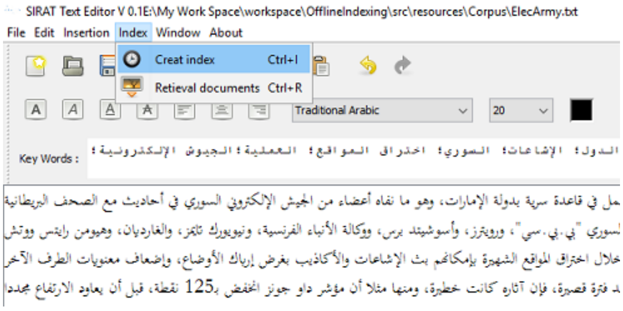
5 Implemented applications

To implement the online semi-automatic indexing system that we designed, we developed an Arabic text editor that contains an online document indexing system. In addition, we worked on building a suitable new form of Arabic corpus, which contains keywords proposed by a human expert, to conduct the necessary experiments. We also used OIRDA application for general indexing and information retrieval and equipped by an offline automatic indexing system of building and updating general index.

5.1 Arabic text editor

We first developed an Arabic text editor (Figure 6), which -in addition to the regular functions as text editor-, is provided with the automatic indexing option to editor's users. We have adopted the design of online semi-automatic indexing system described above (Figure 1) to add this option.

As discussed above, we deal differently with the first document added to the corpus, where there are no other documents, so it only counts a $tf_{i,j}$ value. We then integrate the keyword extraction unit, which is based on the results obtained from the automatic indexing unit prompting some keywords suggestions to the expert indexers, giving them the opportunity to modify these proposed words. Finally, the index is updated. The output of this editor is an object file that contains the processed text and the generated partial index.


File Edit Insertion Index Window About
Creat index Ctrl+I
Retieval documents Ctrl+R
Traditional Arabic
20
Key Words :

SIRAT Text Editor V 0.1E:\My Work Space\workspace\OfflineIndexing\src\resources\Corpus\ElecArmy.txt

Key Words : الدول؛ الإشاعات؛ السوري؛ اختراق المواقع؛ العملية؛ الجيوش الإلكترونية؛

مل في قاعدة سرية بدولة الإمارات، وهو ما نفاه أعضاء من الجيش الإلكتروني السوري في أحاديث مع الصحف البريطانية
لسوري "بي.بي.سي"، ورويترز، وأسوشيتد برس، ووكالة الأنباء الفرنسية، ونيويورك تايمز، والغارديان، وهيومن رايتس ووتش
خلال اختراق المواقع الشهيرة بإمكانهم بث الإشاعات والأكاذيب بغرض إرباك الأوضاع، وإضعاف معنويات الطرف الآخر
د فترة قصيرة، فإن آثاره كانت خطيرة، ومنها مثلا أن مؤشر داو جونز انخفض بـ125 نقطة، قبل أن يعاود الارتفاع مجددا

Figure 6: Arabic Text Editor "SIRAT".

5.2 New Arabic corpus form

To study the efficiency of the proposed system, it was necessary to obtain a test corpus consisting of a set of Arabic documents that would meet a set of necessary and sufficient features for testing.

We have developed a program to build an Arabic corpus, through the organization of a number of web pages of Al Jazeera's website[3], in a new corpus form that is different from the usual ones, by appending keywords suggested by the human expert (Al Jazeera journalists) to the end of documents (Figure 7). This allows evaluating the performance of the automatic keywords extraction unit. In addition, we have taken into account the set of rules used globally in the building of such corpus, especially those provided by (TREC) [34].

```
<DOC>
<DOCNO> ALJAZEERA-511 </DOCNO>
<HEADLINE> ماذا تعرف عن الجيوش الإلكترونية؟ </HEADLINE>
<DATELINE> المصدر: الجزيرة + وكالات, مواقع إلكترونية </DATELINE>
<TEXT>
الجيوش الإلكترونية مجموعات مدربة تعمل وفق أجندة خاصة هدفها اختراق مواقع
الخصوم، والترويج لوجهة نظر معينة عبر مختلف منصات الإنترنت، وإسكات وتشويه
سمعة المناوئين، إلى جانب ترويج الإشاعات والأكاذيب وخلق البلبلة، وقد بدأت الدول
في إنشاء وحدات إلكترونية داخل أجهزتها العسكرية والأمنية لحماية أمنها القومي. خدمة
رسمية الجيوش الإلكترونية مجموعة من الأشخاص وقراصنة الإنترنت (هاكرز) تعمل
لصالح أجهزة المخابرات والأمن في الغالب، تسعى لاختراق المواقع الإلكترونية الخاصة
بالشخصيات والمؤسسات والدول، ولا تكاد تترك منتديات أو نقاشات أو تعليقات على
مواقع التواصل الاجتماعي وغيرها من المواقع الإلكترونية إلا ودخلت إليها للدفاع عن
وجهة النظر الرسمية، ونشر الإشاعات والأكاذيب التي تربك رؤية الناس وتوجههم باتجاه
معين.
…
وفي عصر المعلومة والحروب الإلكترونية، بدأت دول عديدة سياسة إنشاء "جيوش
إلكترونية" نظامية لها ميزانيتها الخاصة، وتسعى للدفاع عن البلاد ضد الهجمات
الإلكترونية التي لا تكاد تنتهي حتى تبدأ. وقد أعلنت وزيرة الدفاع الألمانية أورسولا فون
دير لاين يوم 6 أبريل/نيسان 2017 عن تكوين جيش إلكتروني كوحدة مستقلة داخل
الجيش الألماني إلى جانب القوات البرية والبحرية والجوية، حيث يمارس مهام دفاعية
وهجومية على شبكة الإنترنت. وقالت الوزيرة إن عمل الجيش الإلكتروني لن يقتصر
على صد هجمات القرصنة، بل سيرد عليها أيضا في ساحة المعركة، وهي الإنترنت.
وأضافت "في حال تعرض شبكات الجيش الألماني للهجوم فمن حقنا أيضا أن نرد".
</TEXT>
<KEYWORDS> الجيش الإلكتروني؛ إنترنت؛ حروب إلكترونية؛ فيروسات؛
موسوعة الجزيرة؛ الجيش الإلكتروني السوري </KEYWORDS>
</DOC>
```

Figure 7: New Arabic corpus form.

Thus we were able to obtain an Arabic corpus containing 2416 documents and 25 requests. The vocabulary number of this corpus is 1475148 words, of which 133474 different words (i.e. 9.03% of the total words).

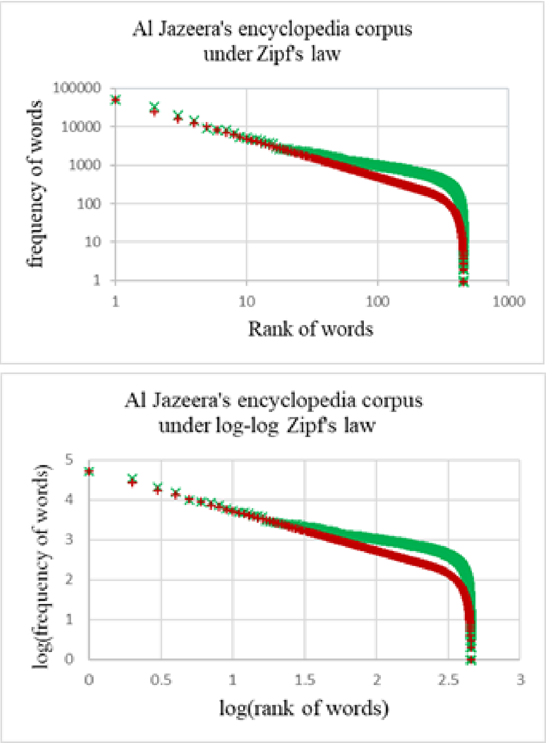


Figure 8: Curve Al Jazeera site corpus according to Zipf's law curve.

According to Zipf's Law, which is concerned with the distribution of words across documents, the range, and highlight the importance of the corpus words. Figure 8 illustrates the Al Jazeera's corpus curve (in red color represented by symbols (+)) and Zipf's curve (in green color represented by symbols (x)). The Figure suggests that Al Jazeera's corpus curve is very close to Zipf's curve. Furthermore, according to some other criteria [31], our new form corpus is very rich and qualified to use as a test collection for IR system quality.

This new format enables us to benefit from, among other things:

- The Contribution to building a system for IRSs evaluation, which enables researchers to test the effectiveness of their applications. In addition to the quality and quantity of the documents considered in this corpus, we have created two types of requests set and their relevant documents. The first is a brief and simple; while the second is extensive and complex, based on the corpus keywords, for example: " الحرب الالكترونية التي يقودها الجيش السوري" (Electronic warfare led by the Syrian Army).
- The Contribution to building a system for keywords systems evaluation, where we have been able to perform extracting experiments using the corpus documents, compare the results of these systems

[3] http://www.aljazeera.net/encyclopedia. Uploaded on November 16, 2017.

with the available keywords and calculate the precision and recall scores.

6 Analysis and results

The aim of our experiments is to evaluate different methods of indexing performance in Arabic information retrieval. A series of experiments was conducted to show the effect of each method of indexing in retrieval performance.

We conducted several experiments using the OIRDA application and endowed it with an offline indexing system for general indexing and information retrieval.

We first compare the following two indexing models:

- *Keyword-based indexing*: the index is composed only of keywords approved by the expert.
- *Indexing without keyword-based or normal indexing*: the index is generated by automatic indexing unit without the intervention of the expert.

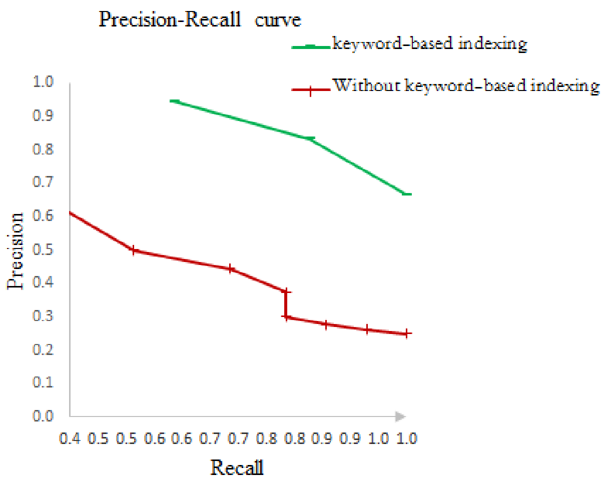


Figure 9: Experiment 1.

Figure 9 represents a comparison between these two models based on their recall-precision curves. The results show that the model of keyword-based indexing, curve in red color represented by symbols (-), is more efficient than the model of indexing without keyword-based, curve in green color represented by symbols (+), on all points of recall and precision.

Then, we compare the two following models of indexing

- *Hybrid*: different combinations of keyword-based indexing and indexing without keyword-based, in a way they token the advantages from each of them.
- *keyword-based indexing*.

In the series of our experiments, the results show that the keyword-based model, curve in green color represented by symbols (-), is more efficient than hybrid model, curve in red color represented by symbols (+). One can observe this behavior in (Figure 10); the curve keyword-based indexing representing the precision based on points of recall is above the hybrid curve.

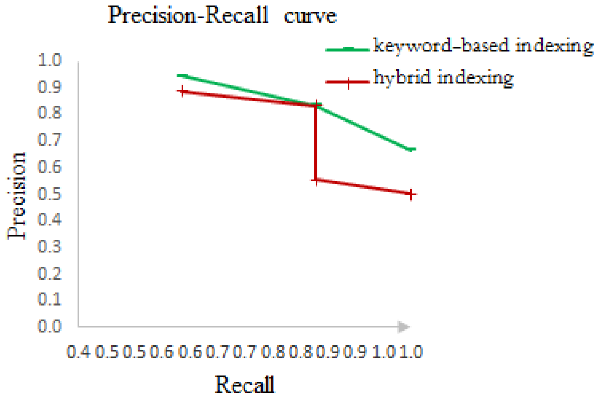


Figure 10: Experiment 2.

The results, further, show that the keyword-based indexing model is the best approach as it is more successful in identifying the descriptors relevant the most to the document. This is primarily due to the intervention of the human expert in keywords identification, especially with ambiguous queries that include polysemy, compound words, etc., which are in need for an accurate semantic processing.

In addition, this model also proved effective to help minimizing index storage size, and thus, improving the response time of the different requests.

The keyword-based indexing model suffers from problems, especially in the case where the expert cannot identify the descriptors that are relevant the most to the document, the aspect this model must improve and find a viable solution to.

7 Conclusion

The main objective of this study is to show the effects of online indexing, which require the semi-automatic indexing, on information retrieval system performance. In addition, this model proved to be effective to help minimize index storage size, and thus, improving the response time of different requests. Therefore, we recommend integrating this model into word processing tools in order to allow the editor to contribute effectively to build a high quality indexes while accounting for the drawbacks and shortcomings of this model. This study also proposes a solution to problems and deficiencies that Arabic language processing suffers from, especially regarding corpus building by developing an application framework for the building and development of corpora. In addition, the paper suggests a solution to reduce deficiencies information retrieval evaluation systems suffer from, which enable researchers to test their indexing and retrieval algorithms and complete systems on common tasks and datasets.

References

[1] S. Bessou, A. Saadi, and M. Touahria, “Un système d’indexation et de recherche des textes en arabe (SITRA).” 1er séminaire national sur le langage naturel et l’intelligence artificielle (LANIA),

Université HAssiba ben Bouali, Département d'Informatique, Chlef (DZ), 2007.

[2] N. Mansour, R. A. Haraty, W. Daher, and M. Houri, "An auto-indexing method for Arabic text," Inf. Process. Manag., vol. 44, no. 4, pp. 1538–1545, 2008. https://doi.org/10.1016/j.ipm.2007.12.007

[3] A. Al Molijy, I. Hmeidi, and I. Alsmadi, "Indexing of Arabic documents automatically based on lexical analysis," Int. J. Nat. Lang. Comput., vol. 1, no. 1, pp. 1–8, 2012.

[4] S. Finch, "Finding structure in language." University of Edinburgh, 1993.

[5] R. Steele and D. Powers, "Evolution and evaluation of document retrieval queries," in Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, 1998, pp. 163–164. https://doi.org/10.3115/1603899.1603927

[6] D. M. W. Powers, "Applications and explanations of Zipf's law," in Proceedings of the joint conferences on new methods in language processing and computational natural language learning, 1998, pp. 151–160. https://doi.org/10.3115/1603899.1603924

[7] J. Entwisle and D. M. W. Powers, "The present use of statistics in the evaluation of NLP parsers," in Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning, 1998, pp. 215–224. https://doi.org/10.3115/1603899.1603935

[8] R. El-Khoribi and M. Ismael, "An intelligent system based on statistical learning for searching in arabic text," ICGST Int. J. Artif. Intell. Mach. Learn. AIML, vol. 6, pp. 41–47, 2006.

[9] L. Khreisat, "Arabic text classification using N-gram frequency statistics a comparative study," Conf. Data Mining| DMIN'06, vol. 2006, pp. 78–82, 2006.

[10] A. M. El-Halees, "Arabic text classification using maximum entropy," IUG J. Nat. Stud., vol. 15, no. 1, 2015.

[11] F. Thabtah, "VSMs with K-Nearest Neighbour to categorise Arabic text data," Proc. World Congr. Eng. Comput. Sci., no. WCECS 2008, October 22-24, 2008, San Francisco, USA, pp. 22–25, 2008.

[12] S. Al-Harbi, A. Almuhareb, and A. Al-Thubaity, "Automatic Arabic text classification," 9es Journées Int. Anal. Stat. des Données Textuelles, pp. 77–84, 2008.

[13] F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, "Naïve Bayesian based on Chi Square to categorize Arabic data," in proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, 2009, pp. 4–6.

[14] T. F. Gharib, M. B. Habib, and Z. T. Fayed, "Arabic Text Classification Using Support Vector Machines," Int. J. Comput. Their Appl., vol. 16, no. 4, pp. 192–199, 2009.

[15] R. Al-Shalabi, G. Kanaan, and M. H. Gharaibeh, "Arabic Text Categorization Using kNN Algorithm," in Proceedings of The 4th International Multiconference on Computer Science and Information Technology, 2006, vol. 4, pp. 5–7.

[16] S. Raheel and J. Dichy, "An empirical study on the feature's type effect on the automatic classification of Arabic documents," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2010, vol. 6008 LNCS, pp. 673–686.

[17] E. Bazzi, M. Salim, T. Zaki, D. Mammass, and A. Ennaji, "Indexation automatique des textes arabes: état de l'art.," E-Ti E-Review Technol. Inf., no. 9, 2016.

[18] N. Tazit, S. S. El Hossin Bouyakhf, A. Yousfi, and K. Bouzouba, "Semantic internet search engine with focus on Arabic language," 2007.

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

[20] M. A. Abderrahim, M. E. A. Abderrahim, and M. A. Chikh, "Using Arabic wordnet for semantic indexation in information retrieval system," arXiv Prepr. arXiv1306.2499, 2013.

[21] A. Mahgoub, M. Rashwan, H. Raafat, M. Zahran, and M. Fayek, "Semantic query expansion for Arabic information retrieval," in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 87–92. https://doi.org/10.3115/v1/W14-3611

[22] F. Harrag, E. El-Qawasmah, and A. M. S. Al-Salman, "Stemming as a feature reduction technique for Arabic text categorization," in Proceedings of the 10th International Symposium on Programming and Systems, ISPS' 2011, 2011, pp. 128–133. https://doi.org/10.1109/ISPS.2011.5898874

[23] R. Mohamed and J. Watada, "An evidential reasoning based LSA approach to document classification for knowledge acquisition," in IEEM2010 - IEEE International Conference on Industrial Engineering and Engineering Management, 2010, pp. 1092–1096. https://doi.org/10.1109/IEEM.2010.5674188

[24] F. S. Al-Anzi and D. AbuZeina, "Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing," J. King Saud Univ. Inf. Sci., vol. 29, no. 2, pp. 189–195, 2017. https://doi.org/10.1016/j.jksuci.2016.04.001

[25] T. Dilekh and A. Behloul, "Implementation of a New Hybrid Method for Stemming of Arabic Text," Analysis, vol. 46, no. 8, pp. 14–19, 2012.

[26] S. R. El-Beltagy and A. Rafea, "KP-Miner: A keyphrase extraction system for English and Arabic documents," Inf. Syst., vol. 34, no. 1, pp. 132–144, 2009. https://doi.org/10.1016/j.is.2008.05.002

[27] E. Amer and K. Foad, “Akea: an Arabic keyphrase extraction algorithm,” in International Conference on Advanced Intelligent Systems and Informatics, 2016, pp. 137–146.

[28] M. Al-Kabi, H. Al-Belaili, B. Abul-Huda, and A. Wahbeh, “Keyword extraction based on word co-occurrence statistical information for arabic text,” Abhath Al-Yarmouk" Basic Sci. Eng., vol. 22, no. 1, pp. 75–95, 2013.

[29] T. El-Shishtawy and A. Al-sammak, “Arabic Keyphrase Extraction using Linguistic knowledge and Machine Learning Techniques,” ReCALL, pp. 1–8, 2012.

[30] R. Duwairi and M. Hedaya, “Automatic keyphrase extraction for Arabic news documents based on KEA system,” J. Intell. Fuzzy Syst., vol. 30, no. 4, pp. 2101–2110, 2016. https://doi.org/10.3233/IFS-151923

[31] Y. Kadri and J. Y. Nie, “Effective stemming for Arabic information retrieval,” Proc. Chall. Arab. nLP/mt, Int. conf. Br. Comput. Soc., pp. 68–74, 2006.

[32] I. A. Al-Kharashi and M. W. Evens, “Comparing words, stems, and roots as index terms in an Arabic information retrieval system,” J. Am. Soc. Inf. Sci., vol. 45, no. 8, p. 548, 1994. https://doi.org/10.1002/(SICI)1097-4571(199409)45:8<548::AID-ASI3>3.0.CO;2-X

[33] K. Beesley, “Arabic morphological analysis on the Internet,” in Proceedings of the 6th International Conference and Exhibition on Multi-lingual Computing, 1998.

[34] L. S. Larkey and M. E. Connell, “Arabic Information Retrieval at UMass in TREC-10.,” in TREC, 2001.

[35] E. M. Voorhees, “Overview of TREC 2003.,” in Trec, 2003, pp. 1–13.