Query Expansion for Arabic Information Retrieval Model: Performance Analysis and Modification

Ayat Elnahaas[*1], Nawal Alfishawy[**2], Mohamed Nour [*3], Gamal Attiya[**4], Maha Tolba[**5]

[*]*Department of Research Informatics, Electronics Research Institute, Cairo, Egypt*

[1]eng_ayatelnahas@yahoo.com; [3]mnour@eri.sci.eg

[**]*Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University, Egypt*

[2]nelfishawy@hotmail.com; [4]gamal.attiya@yahoo.com; [5]maha_saad_tolba@yahoo.com

Abstract- ***Information retrieval aims to find all relevant documents responding to a query from textual data. A good information retrieval system should retrieve only those documents that satisfy the user query. Although several models were developed, most of Arabic information retrieval models do not satisfy the user needs. This is because the Arabic language is more powerful and has complex morphology as well as high polysemy. This paper first investigates the most recent Arabic information retrieval model and then presents two different approaches to enhance the effectiveness of the adopted model. The main idea of the proposed approaches is to modify and/or expand the user query. The first approach expands user query by using semantics of words according to an Arabic dictionary. The second approach modifies and/or expands user query by adding some useful information from the pseudo relevance feedback. In other words, the query is modified by selecting relevant textual keywords for expanding the query and weeding out the non-related textual words. The adopted retrieval model and the two proposed approaches are implemented, tested, compared, and evaluated considering Arabic document collection. The obtained results show that the proposed approaches enhance the effectiveness of the Arabic information retrieval model by about 15% to 35%.***

Keywords: ***Arabic Documents, Indexing, Vector Space Model, Query Expansion, Semantics, and Relevance Feedback.***

1 INTRODUCTION

Information retrieval is one of the most important research areas in information technology. The main objective is to match and retrieve the most relevant documents to the user query. Therefore, a good information retrieval system should retrieve only those documents that satisfy the user needs.

Generally, an information retrieval system contains several modules mainly: document collection, query processing, matching operations and query performance [1]. Figure 1 shows the main modules of an information retrieval system [2]. Document collection and representation involves an important process called indexing. The indexing process associates a document with a descriptor represented by a set of features automatically derived from the content. It also optimizes the query performance and improves the response time by sorting terms in an interested file structure. Moreover, a number of processing tasks can take place during the indexing phase similar to the query processing which further improves the performance [3-5]. Document-query matching aims to estimate the relevance of a document to the given query. Most information retrieval models compute a relevance score. This score is used as a criterion to rank the list of documents retrieved to the user in response to the query. That is, the results of matching between the user query and the index terms are posted based on a Ranking method.

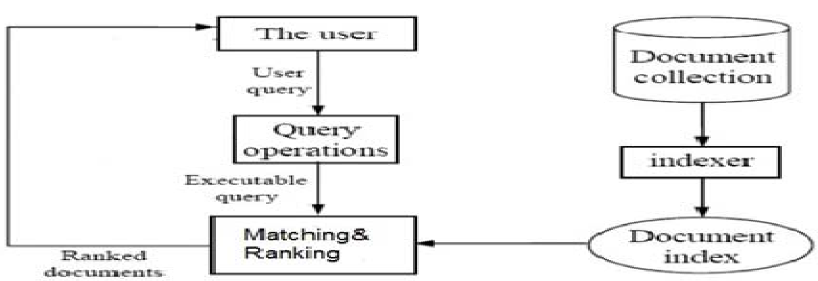


Figure1: The Main Modules of an Information Retrieval System [2]

A natural language query specifies the user's information need in a sentence. Representing the user need involves query formulation using terms expressed by the user and/or additive information driven by iterative query improvements like relevance feedback. The query/ user need is parsed and compiled into an international form. In case of textual retrieval, query terms are generally preprocessed to select the index objects. The query representation involves one-step or multi-step query formulation driven by iterative query improvements [6-8]. The querying stage involves many themes including query preprocessing, removal of stop-words, query expansion, and others. The query expansion expands the query with similar terms and then retrieves another set of documents using expanded query [6, 9]. Moreover, information retrieval implements a basic term matching for identical terms. The document-query matching is known as query evaluation for estimating the relevance of documents to the given query. The information retrieval system employs some ranking methods based on mathematical bases to exploit some properties found in the document collection. Matching between the query keywords and index terms may be exact matching, partial matching, or intelligent matching [10].

Although several models were developed [11-17], most of Arabic information retrieval models do not satisfy the user needs. This is because the Arabic language is known with its powerful and complex morphology as well as its high polysemy. This paper first investigates an Arabic information retrieval model and then presents two different approaches to enhance the effectiveness of the model. The focus is concerned with the modification and/or expansion of the user query. The first approach expands user query by using semantics of words according to an Arabic dictionary. The second approach modifies and/or expands user query by adding some useful information from the pseudo relevance feedback. In other words, the query is modified by selecting relevant textual keywords for expanding the query and weeding out the non-related textual words. The proposed approaches are implemented, tested and evaluated using some measurable criteria such as precision, recall, and F-measure. In addition, the obtained results are compared with that obtained by the most recent adopted Arabic retrieval model for Arabic document collection [2].

The rest of this paper is organized as follows. Section 2 presents a literature survey for related work. Section 3 presents an adopted information retrieval model. Section 4 presents the proposed approaches and describes the query expansion using semantics of keywords and relevance feedback. Section 5 presents the simulation experimental results and discussions while section 6 concludes this work.

2 RELATED WORK

Regarding the information retrieval systems/ models, several research efforts were presented by a lot of researchers [11-17]. In [11], the authors mentioned that any information retrieval model can be represented by four attributes: D, Q, F, and R. D is the set of documents in the document collection. Q is the set of queries representing the user needs. F is concerned with classical document representation, queries, and their relationships. R is a ranking function $R(q_i,d_j)$ which affiliates a real number with a query q_i € Q and a document representation d_j € D. In [12], the authors mentioned that the Boolean model is one of the oldest information retrieval models. That model uses the set theory and/or Boolean algebra. The user Query can be represented by a set of keywords connected together logically by a set of connections like AND, OR, and NOT. The AND operator produces the set of documents of both sets. The OR operator produces a document set that is bigger than or equal to the document sets of any of the single terms. The NOT operator is used to avoid retrieving a document containing a specific keyword. In [13], the authors discussed the vector space model that represents the documents and queries as vectors in a multidimensional space. To assign a numeric score to a document for a query, the model measures the similarity between the query vector and the document vector. The angle between two vectors is used as a measure of divergence between the vectors. The cosine angle is used as the numerical similarity. If the cosine angle has the value '1' it means the vectors are identical while the vectors are orthogonal if the cosine angle has the value '0'. The vector space model is good as it attempts to rank documents by some similarity values between the user query and each document. In [14, 15], the authors discussed the probabilistic model of information retrieval which relies on the notion that each document has a certain probability of being relevant to a query. The documents that are most likely to be relevant and useful to the user are ranked by a decreasing order of probability. For two events A and B, the joint event of both events occurring is described by the joint probability P(A, B). The conditional probability P(A|B) expresses the probability of event A given that B occurred. Probabilistic information retrieval models include classic Probabilistic models, language models and the relevance model. All those models have variants that incorporate word dependence.

In [16], the authors conducted the process of developing ontology for Arabic Blogs retrieval. The authors mentioned that semantic search engines provide searching and retrieving resources related to the user's need. The authors proposed a model for representing Arabic knowledge in the computer technology domain using ontologies. The model was concerned with elicitation of user's information needs. Ontologies play a vital role in supporting information search and retrieval process of Arabic blogs on the web. In [17], the authors presented an enhanced Arabic information retrieval approach. The focus was on the effectiveness of using the list of stop-words and light stemming of Arabic. The authors used the vector space model as a popular weighting scheme in their work. Their work aims at combining the stop-words list with light stemming to enhance the performance and compare their effects on retrieval. The authors tested their adopted approach using the Arabic news consortium dataset. In [9], the authors discussed the concept of query expansion for improving the process of Arabic information retrial. The query expansion was based on the similarity of terms. The authors employed the expectation-maximization algorithm for selecting the relevant terms and weeding out the non-relevant ones. They tested performance of the adopted algorithm using INFILE test collection. The experiments indicate good performance of precision and recall for the used query expansion method.

3 ADOPTED ARABIC INFORMATION RETRIEVAL MODEL

In 2016, an adopted Arabic information retrieval is developed [2]. The authors discussed the main challenges of Arabic query expansion using Word-Net and association rules. They mentioned that they are able to exploit Arabic word-Net to improve the retrieval performance. Their obtained results on a sub-corpus from the Xinhua collection showed that the automatic selection method is significant and improves the performance of information retrieval systems. The adopted Arabic information retrieval model [2] involves important themes mainly: preprocessing, document collection and indexing, user query, and matching operations.

A. Preprocessing

The preprocessing steps are done on the document terms before building the index and on the user query before matching process. The preprocessing should be done first to gain the benefit of speeding-up the retrieval time [18, 19]. The preprocessing steps involve tokenization, removal of stop-words and stemming.

1) Tokenization

Tokenization; in natural language processing; means splitting text into tokens. A token is the smallest unit of text that may be a word, a punctuation mark or a multi-word expression. The separator between two adjacent words may be a white space or punctuation marks. Tokenization is an important step for most natural language processing tasks [37]. In this work, Lucene Arabic tokenizer is used during the implementation of this stage http://www.apache.org/licenses/LICENSE-2.0. Figure 2 shows an example of a document title before and after tokenization.

أهم المركبات المسموح بها في الزراعة العضوية لمقاومة الأمراض والحشرات

(a) A document title before tokenization

أهم,المركبات,المسموح,بها,في,الزراعة,العضوية,لمقاومة,الأمراض,و,الحشرات

(b) The document title after tokenization

Figure 2: A document title tokenization

2) Removal of Stop-words

Removal of stop-words means rejecting the useless words like preposition, pronoun, specifiers, modifiers, and other tools. Examples of the stop words are: - الذي، هي، هو، في، علي، إلي، من . Such words frequently occur in Arabic documents. These words don't give any hint for the content of their documents. In information retrieval systems, stop-words should be eliminated (by referring to a stop-word list) from the query text and from the set of index terms [18, 20]. Figure 3 shows the tokens of a document title after removing the stop-words.

أهم، المركبات، المسموح، الزراعه، العضويه، مقاومة، الأمراض، الحشرات

Figure 3: The tokens of the document title, in Figure 1, after removal of stop-words

3) Stemming

The stemming process is very important for Arabic information retrieval. Stemming aims at reducing all of the inflectional derivational variants of words into a common form called the stem. A word stem can be obtained by removing all the affixes attached to the word. The words sharing some root or stem can increase the matching of documents to the user query. Stemming can reduce the index size and improve the performance of the retrieval process. Figure 4 shows the tokens of a document title after stemming.

أهم، مركب، مسموح، زراع، العضوي، مقاوم، الأمراض، الحشرات

Figure 4: A Document Title after Stemming

There are several types of stemmers. Examples of Arabic stemmers are: light stemmer (light 10), Khoja stemmer, Porter stemmer, and others. In this work, Porter stemmer is used during the implementation of this stage [18, 21-23]. For more details about the Porter stemmer mechanism, the reader can refer to the website https://tortous.org/mortim/porter.stemmer.

B. Document Collection and Indexing

Indexing is the process of choosing a term or a number of terms that can represent what the document contains. In other words, after doing the preprocessing steps on the chosen document collection, the index can be built. Each document is represented by a set of important terms, which were taken from the document title. Such terms are weighted and stored in an index (as index terms) without any repetition. The index contains document number, terms, frequency/weight in addition to other useful information such as the number of documents that contain each term. Figure 5 shows an Arabic example of a part of the index mapping [24-26]. The index terms will be matched against the query keywords.

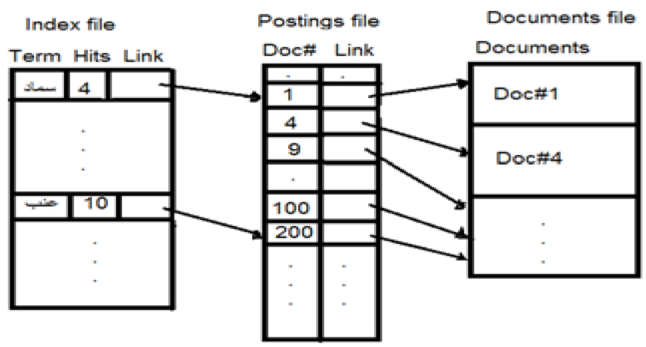


Figure 5: Arabic Example of a Part of the Index Mapping

C. User Query

The querying stage is handled exactly like the document. That is, the preprocessing steps; tokenization, removal of stop-words, and stemming are done on the input user query. The user query may be a word, phrase, or sentence containing a set of keywords. If the query is one word, the stemming operation only can be done. If the query contains a set of words, it should be preprocessed (tokenization, stop-words removal, and stemming). In this work, several queries are presented and processed. Some of the queries contain only one keyword while others contain two keywords, three keywords, and four keywords respectively. Table1 shows some examples of the user independent queries while Table 2 contains examples of some related queries.

TABLE1: EXAMLPES OF USER INDEPENDET QUERIES

User Query	No. of Keywords
التين	1
زراعة الخضروات	2
صادرات مصر من القمح	3
أهمية البلح وطرق تجفيفة	4

TABLE 2:EXAMPLES OF SOME RELATED QUERIES

User Query	No. of Keywords
العنب	1
محصول العنب	2
الجديد فى محصول العنب	3
ألجديد فى إنتاج محصول العنب	4

D. Matching and Ranking

The matching process is done between the query keywords and document terms. To facilitate the matching process, a matching model is used. In this paper, the Vector Space Model (VSM) is used for the matching operation [18, 20, 27-32].
The VSM is an algebraic model where it uses non-binary weights that are assigned to the index terms of documents and queries. The document set D is represented as follows:-

$$D = \{d_1, d_2, d_3 \ldots d_N\} \quad (1)$$

where, d_j is the document number j, and N is the number of documents in the dataset collection.
Any document d_j is represented by a set of terms' weights as follows: -

$$d_j = \{w_{1j}, w_{2j}, w_{3j} \ldots w_{mj}\} \quad (2)$$

where, w_{1j} is the weight of the term i in the document j. The weight of term i in document j can be calculated using the term frequency (tf) and inverse document frequency (idf). So,

$$w_{ij} = tf_{ij} * idf_i \quad (3)$$

where, the term frequency tf_{ij} is the number of occurrence of term i in the document j and idf_i is the inverse document frequency of term i.

$$idf_j = \log_2 \frac{N}{n_i} \quad (4)$$

where, n_i is the total number of occurrence of item i in all documents.
Documents can be retrieved and ranked by matching the query vector versus the document vector to compute the score or similarity. The retrieved documents are ranked according to the similarity to the user query [33-36].

$$sim(d_j q_i) = \frac{\sum_{i=1}^{t} w_{ij} * w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2} * \sqrt{\sum_{i=1}^{t} w_{iq}^2}} \quad (5)$$

where, $sim(d_j, q_i)$ is the similarity between document j and query q_j, w_{ij} is the weight of term i in document j, and w_{iq} is the weight of term i in query q.

4 PROPOSED APPROACHES

This section presents two new efficient approaches to enhance the effectiveness of the most recent adopted Arabic information retrieval model [2]. The main idea of the proposed approaches is to modify and/or expand the user query by using semantics of words in the first approach and using some useful information from the pseudo relevance feedback in the second approach.

A. Query Expression using Semantics of Keywords

Query expansion means adding extra new terms to the keywords of the initial query. Since the input user query has the significant effect on the document retrieval, hence the user query may be modified and/or expanded to retrieve more relevant documents. The addition of new terms should take place prior the initial search.
It is known that Arabic is one the Semitic languages. Arabic has a rich set of vocabularies. Arabic language is polysemous as the same word may have several meanings. Moreover, the Arabic language has a different morphological structure for its wide range of derivations [2, 4]. By searching the dictionary for the meaning of an Arabic keyword, more than one meaning may be found. This is the case for the majority of Arabic words. This means that each query keyword has multiple synonyms/meanings.
In this paper, the first proposed approach expands user query by using semantics of words. In this case, the synonyms or semantics of the query keywords can be obtained by referring to either the Arabic Word-Net or Arabic dictionary. In the first

approach, semantics of the query keywords are chosen according to an Arabic dictionary. Expanding the query to include more or extra keywords will improve the performance of the retrieval model as it presents more relevant documents to the user.

To illustrate the query expansion method, let Q be the set of queries entered separately from the user, where Q = $\{q_1, q_2, q_3, \ldots q_r\}$. Each query q_r has a set of m keywords. That is, $q_r = \{k_1, k_2, \ldots k_m\}$, where k_i is the query keywords which represents the user needs and $1 \leq i \leq m$. By searching the dictionary for the meaning of each keyword, a list S_{ki} of n synonyms associated to the keyword k_i may be found, i.e., $S_{ki} = \{S_{i1}, S_{i2}, S_{i3}, \ldots, S_{in}\}$. Each list S_{ki} contains the number of synonyms associated to a keyword k_i in the query and $1 \leq i \leq m$. This means that the number of synonyms' lists of a query q_r equals the number of keywords in the initial use query. That is, $S(q_r) = \{S_{k1}, S_{k2}, \ldots S_{km}\}$, where $S(q_r)$ is the set of lists.
Figure 6 shows the associated synonyms of query keywords. From Figure 6, each query keyword k_i has multiple synonyms/meanings S_{ij} where $1 \leq i \leq m$ and $1 \leq j \leq n$. Moreover, it is not necessary for all query keywords to have the same number of corresponding meanings. For this reason, we focus here on using only one meaning which is the commonly used one. The chosen meaning is taken based on its strong relation with the keyword. That is, the number of query keywords after expansion becomes the double of the original one. To illustrate that concept, some simple examples are given in Table 3. The query expansion can extract the equivalent terms of query keywords from the relation between the concepts or meanings such as:

(زراعة، فلاحة)، (طماطم، بندورة)، (كرنب، ملفوف)، وهكذا.

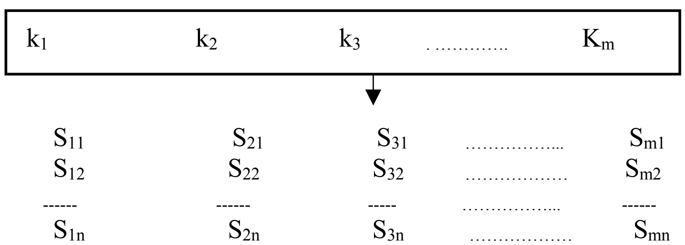


Figure 6: Query Keywords and their Semantics

TABLE 3: USER QUERY EXPANSION USING SYNONYMS/SEMANTICS

Initial Query	Expanded Query
زراعة العنب	زراعة، فلاحة ، العنب ، الكرم
تسميد الكرنب	تسميد، تخصيب ، الكرنب ، الملفوف
إنتاج البلح	إنتاج، البلح، التمر

B. Query Expansion using Relevance Feedback

As mentioned above, query expansion aims to add extra terms or more information to clarify the user query. The query expansion helps in matching more additional documents. In this paper, the second proposed approach modifies and/or expands user query by adding some useful information from the pseudo/user relevance feedback. In other words, the query is modified by selecting relevant textual keywords for expanding the query and weeding out the non-related textual words. The idea is going to keep track of those terms that should be added to the query and those should be eliminated.
The process of query expansion by the principle of user relevance feedback may be described as follows:

1. The original keywords of the user query, after doing the preprocessing operations, are matched against the index terms. The retrieved documents are presented from the highest to lowest values depending on the similarity values.
2. The retrieved documents should be analyzed to monitor and identify their terms' descriptors. This is important to add those terms appeared in the relevant documents to the original user query and also to eliminate those terms describing the retrieved irrelevant documents.
   2.1 A maximum threshold value (max_{th}) of documents similarities should be defined. This means that the terms' descriptors for only those retrieved documents with similarity values $\geq mac_{th}$ will be chosen to be added to the original query keywords.
   Let S_1 be the set that collects all relevant retrieved documents that satisfy the threshold condition max_{th}.

$$S_1 = \{ d_1, d_2, \ldots\ldots\ldots max_{th}\} \quad (6)$$

2.2 A minimum threshold value (min_{th}) of documents similarities is defined. This means that the terms' descriptors for only those retrieved documents with similarity values ≤ min_{th} will be eliminated from the query. Let S_2 be the set that gathers all non-relevant retrieved documents and the min_{th} condition is satisfied

$$S_2=\{ d_1, d_2, \ldots\ldots\ldots d_y\} \tag{7}$$

3. The query can be expanded by adding the terms of the selected relevant documents from S_1 and also eliminating those terms of the chosen irrelevant documents from S_2. That is

$$q_{\exp} = q_{user} + \sum_{di \in s_1} d_i - \sum_{di \in s_2} d_j \tag{8}$$

5 SIMULATION RESULTS AND DISCUSSION

This section presents several experimental to evaluate the performance of the proposed approaches. To do so, the adopted information retrieval model [2] and the proposed approaches are implemented and tested considering a dataset in the agriculture field. The performance is evaluated using some measurable criteria such as precision, recall, and F-measure.

A. Simulation Environment

The proposed approaches are implemented using JAVA programming language besides Lucene APIS, which is a powerful searching library, using an HP-Labtop with a processor 2.5 GHZ, and Windows-7 operating systems. The approaches are coded in JAVA and supported by the Apache software foundation.

B. Document Collection Dataset

To check the efficiency of the proposed approaches against the adopted information retrieval model [2], they are operated and tested using a chosen document collection as a test-bed. The documents in the dataset are acquired from different Arabic websites mainly http://www.kenanaonline.net/page/Agriculture and http://www.zeraiah.net/index.php/baydar. The test-bed documents are in the agriculture field. It contains four hundred documents. Each document has a document title and contents. Each document is represented by a set of important terms, which were taken from the document title. Such terms are weighted and stored in an index (as index terms) without any repetition. The index terms will be matched against the query keywords.

C. Performance Metrics

The performance is evaluated using some measurable criteria such as precision, recall, and F-measure. These criteria are defined as follows [19-20].

$$\Pr ecision = \frac{\text{number of the relevant retrieved documents}}{\text{number of the retrieved documents}} \tag{9}$$

$$\text{Re}\,call = \frac{\text{number of the relevant retrieved documents}}{\text{number of the relevant documents}} \tag{10}$$

$$\text{F - measure} = \frac{2(\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}} \tag{11}$$

D. Experimental Results

Several experiments are done to test and monitor the performance of the adopted information retrieval model and the proposed approaches. Four categories of queries are adopted with five different queries for each. The query categories have one keyword, two keywords, three keywords, and four keywords respectively. The queries in Figures 7, 8, 9, and 10 are independent. The queries in Figure 11 are related to each other, i.e., the keyword of query#1 exists in query#2. The two keywords of query#2 exist in query#3 and the three keywords of query#3 exist in query#4. This is also the case for other queries in Figures 12, and 13 respectively.

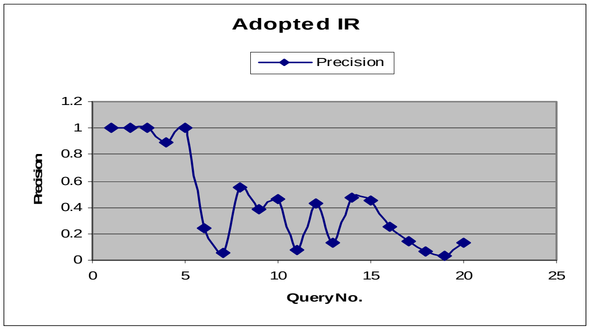


Figure 7a: Precision for Adopted IR

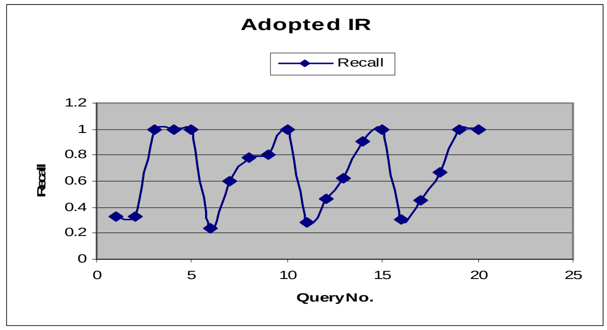


Figure 7b: Recall for Adopted IR

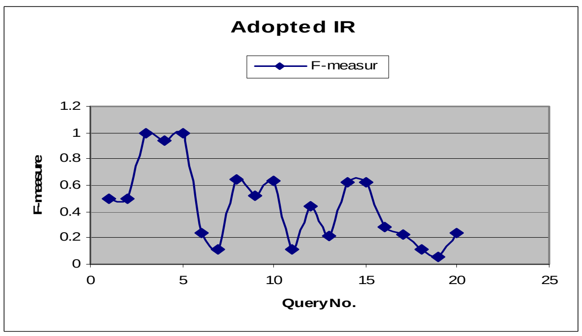


Figure 7c: F-measure for Adopted IR

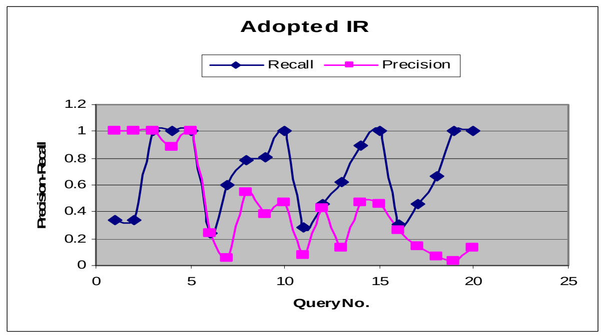


Figure 7d: Precision-Recall for Adopted IR

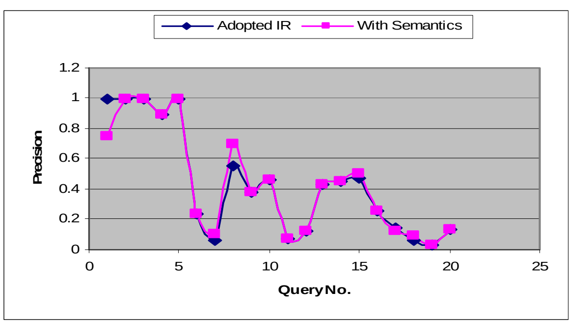


Figure 8a: Adopted IR and Keywords' Semantics

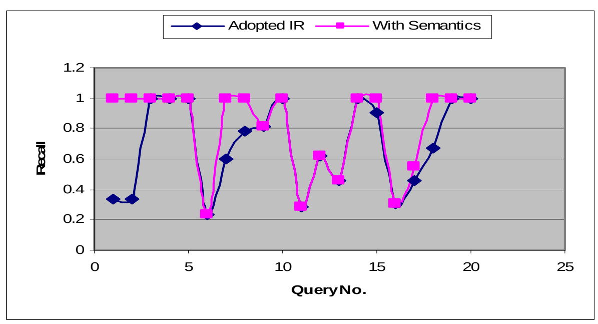
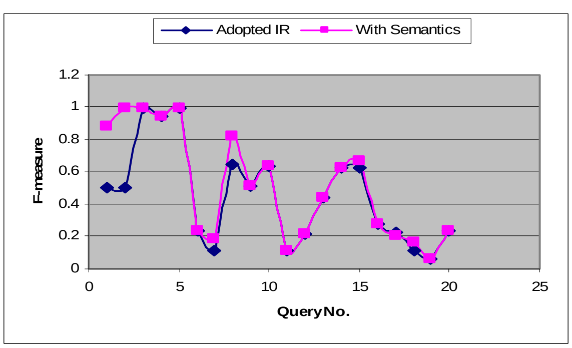


Figure 8b: Adopted IR and Keywords' Semantics



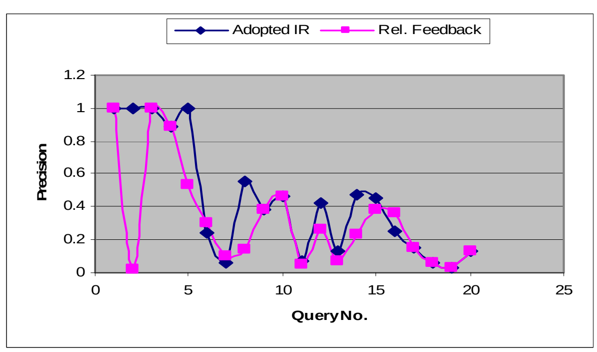
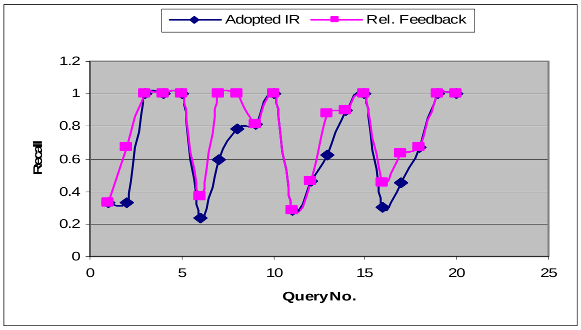


Figure 8c: Adopted IR and Keywords' Semantics

Figure 9a: Adopted IR and Relevance Feedback



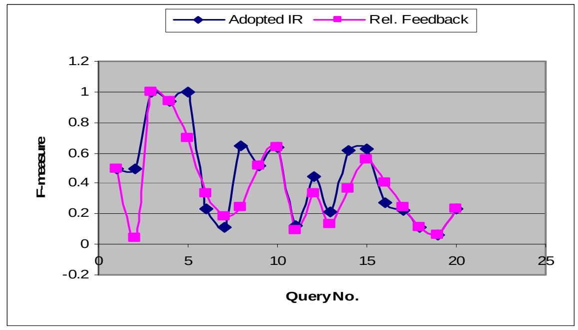
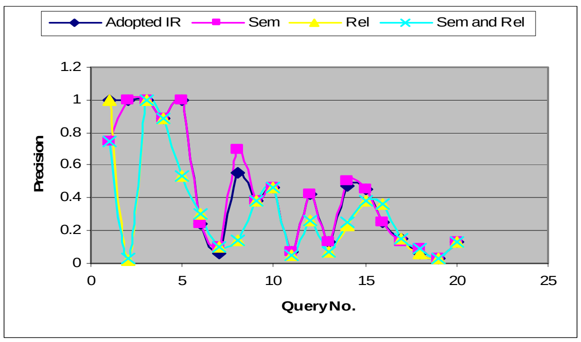


Figure 9b: Adopted IR and Relevance Feedback

Figure 9c: Adopted IR and Relevance Feedback



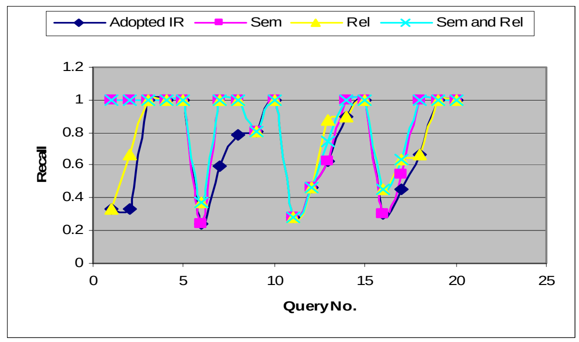
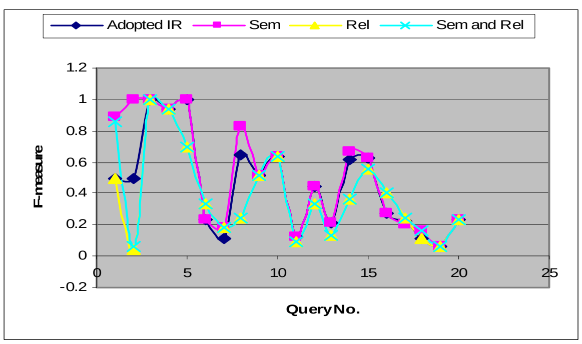


Figure 10a: Adopted IR, Sem, Rel, and Both

Figure 10b: Adopted IR, Sem, Rel, and Both



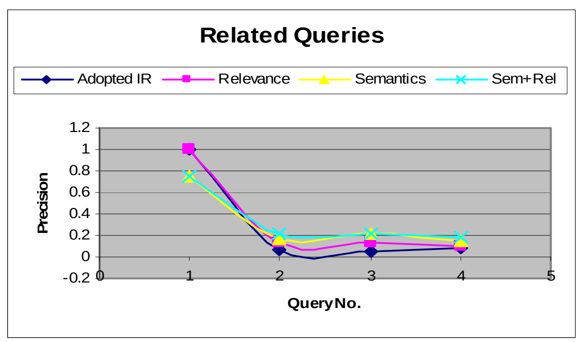
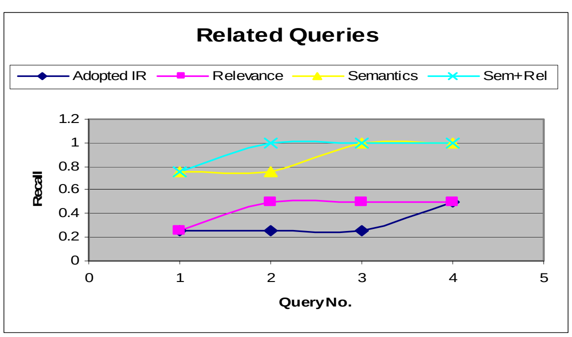


Figure 10c: Adopted IR, Sem, Rel, and Both

Figure 11a: Precision for Related Queries



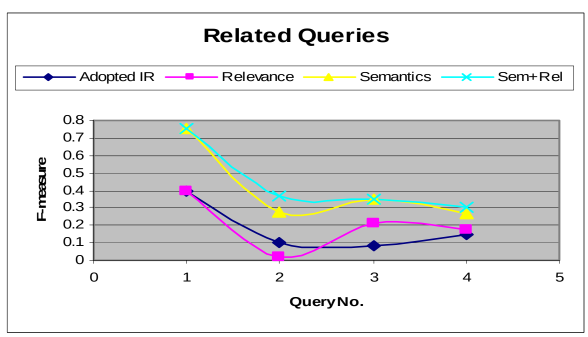


Figure 11b: Recall for Related Queries

Figure 11c: Precision for Related Queries

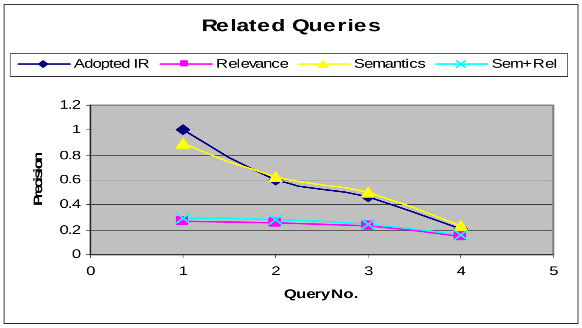


Figure 12a: Recall for Related Queries

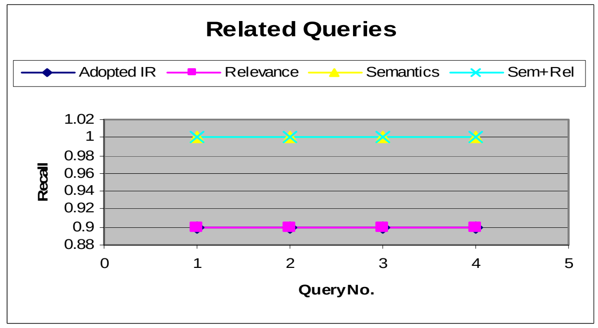


Figure 12b: Precision for Related Queries

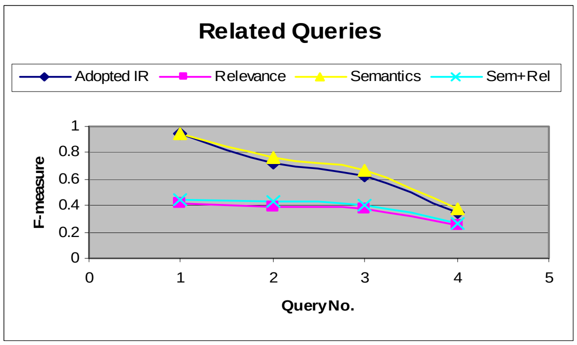


Figure 12c: F-measure for Related Queries

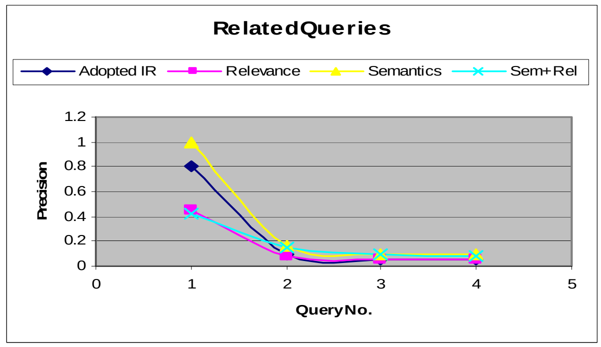


Figure 13a: Precision for Related Queries

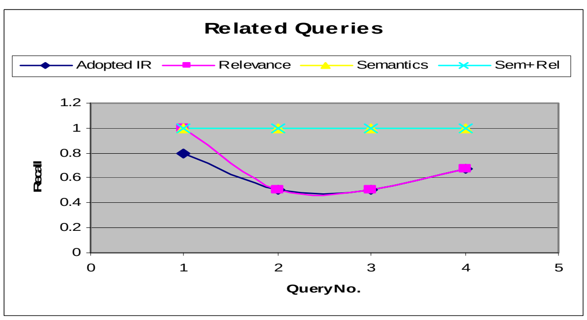


Figure 13b: F-measure for Related Queries

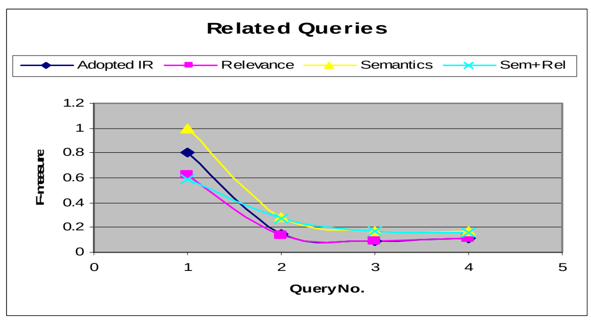


Figure 13c: Precision for Related Queries

From the experimental results, the values of precision, recall, and F-measure for each query are different from those of the other queries either in the same query category or in other categories. This means that the concept type where a query is asking for is significant and this is clear in all experiments. If the number of query keywords changes, the values of precision, recall, and F-measure are also changing. This is clear in Figures 7, 8, 9, and 10 respectively. In Figures 7, 8, 9, and 10, a query has one keyword for the first five queries, two keywords for the second five queries, three keywords for the third five queries, and four keywords for the fourth five queries respectively. This means that the number of query keywords has a direct effect on the retrieval performance. If the precision value changes then the recall and F-measure values are also changing. This happened in the majority of the test-bed queries. We don't have a guarantee to say that increasing values always appear in a linear or a nonlinear form.

The values of precision, recall, and F-measure for the adopted retrieval model with semantics are different from those corresponding values without semantics. This means that the modified approach for the query expansion using semantics of keywords has a positive effect. This is clear in all query categories in Figures 7 and 8 respectively.
A precision value; regardless the query concept and query keywords; is always increasing or in some cases remains fixed compared to the adopted retrieval model without semantics. This is because the number of relevant retrieved documents is

increasing or sometimes remaining unchanged. The values of precision, recall, and F-measure are better for the modified approach than those corresponding values of the adopted retrieval model without semantics. This is clear in Figures 8a, 8b, and 8c respectively. The improvement in performance for the modified approach using semantics of keywords ranges from 8% to 27% depending on the number of query keywords as well as the query concept.

The values of precision, recall, and F-measure for the test-bed queries for the modified approach using relevance feedback are better than their corresponding values of the model without modification. This is clear in Figures 9a, 9b, and 9c respectively. The improvement of performance is slightly better than that model without modification.
The values of precision, recall, and F-measure for the modified approach using both semantics of keywords and relevance feedback are better than those without any modification. This is clear in Figures 10a, 10b, and 10c respectively. The improvement values for the adopted experiments are ranging from 15% to 34% depending on the query concept and query category. Moreover, the performance of the retrieval model is better modified using keywords' semantics than that using only relevance feedback. In other words, combining both the relevance feedback and semantics makes slightly change in precision, recall, and F-measure compared to that one using only semantics of keywords. This is clear in Figures 10a, 10b, and 10c respectively. The improvement values are in the range of 3% to 13%.

Moreover, three different experiments with four related queries per each are also implemented and run as shown in Figures 11, 12, and 13 respectively. The experiments are tested and compared among the performance of the adopted retrieval model and the two modified approaches for query expansion using semantics of keywords, relevance feedback, and both. From the experimental results shown in Figures 11, 12, and 13 respectively, it is shown that the values of precision, recall, and F-measure are better for the modified approaches than those corresponding values of the adopted information retrieval without modification.

6 CONCLUDING REMARKS

In this research work, the most recent adopted information retrieval model was investigated and analyzed. In addition, two new efficient approaches are developed to enhance the effectiveness of the recent model. The adopted model is modified by expanding the queries using semantics of keywords and/or relevance feedback. The models are implemented and tested using an Arabic document collection test-bed. From the practical results, the representation and formulation of a user query plays an important role in the performance of the information retrieval model. The query expansion increases the number of retrieved relevant documents. The obtained results showed that the values of precision, recall, and F-measure for the two modified approaches are better than that without modification. The query expansion using word semantics improve the performance by about 27% compared to the original model. While, the query expansion using relevance feedback improve performance by about 14%. Finally, combining both the semantics of keywords and query relevance feedback for expanding the user queries outperforms the adopted retrieval model without modification. The hybrid query expansion using the two modifications improves the performance by 15% to 35%.

REFERENCES

[1] Ghaith AbdulSattar Alkubaisi, "Design and Implementation of Knowledge-Based System for Text Retrieval Based on Context and User's Prior Knowledge" M.Sc. thesis, Department of Computer Science, Faculty of Information Technology, Middle East University, Amman, 2013.

[2] Ahmed Abbache, Farid Meziane, Ghalem Belalem, and Fatma Bellredim, "Arabic Query Expansion using Word-Net and Association Rules", The International Journal of Intelligent Information Technologies, Vol. 12, No. 3, pp. 51-64, July-September 2016.

[3] Soner Kara, O¨zg¨ ur Alan, Orkunt Sabuncu, Samet Akpınar, Nihan K. Cicekl and Ferda N. Alpaslan, "An Ontology-based Retrieval System Using Semantic Indexing", https://etd.lib.metu.edu.tr/upload/12612110/in dex.pdf., Downloaded in 2016.

[4] Emad Elabd, Eissa Alshai, and Hatem Abdulkader, "Semantic Boolean Arabic Information Retrieval", The International Arab Journal of Information Technology, Vol. 12, No. 3, pp. 311-316, May 2015.

[5] Eissa Mohammed Mohsen Alshari, "Semantic Arabic Information Retrieval Framework", M.Sc. Thesis, Information Systems Department, Faculty of Computers and Information, Menoufiya University, 2014.

[6] Fatiha Boubekeur, and Wassila Azzoug, "Concept-Based Indexing in Text Information Retrieval", The International Journal of Computer Science and Information Technology (IJCSIT), Vol. 5, No. 1, pp. 119-136, Feb. 2013.

[7] Miriam Fernandez, Ivan Cantador, Vanesa Lopez, David Vallet, Pablo Castells, and Enrico Motta, “Semantically Enhanced Information Retrieval: An Ontology-based Approach”, Downloaded in 2017 from http://www.elsevier.com/ locate/websem.

[8] Komal Shivaji Mule, and Arti Waghmare, “Improved Indexing Technique For Information Retrieval Based On Ontological Concepts”, The International Journal of Computer Applications (IJCA) and the National Conference on Advances in Computing (NCAC), pp. 5-20, 2015.

[9] Khaled Shaalan, Sinan Al-Shaikh, and Farhad Oroumchian, “Query Expansion Based on Similarity of Terms for Improving Arabic Information Retrieval”, A Technical Report Presented to the University of Wollongong in Dubai, The 7th IFIP TC12 International Conference on Intelligent Information Processing, Springs, Heidelberg, pp. 167-176, 2012.

[10] Fausto Giunchiglia, Uladzimir Kharkevich, and Llya Zaihrayeu, “Concept Search", Downloaded in 2016 from the http://eprints .biblio.unitn.it/1434/1/037.pdf.

[11] Djoesd Hiemstsa, “Information Retrieval Models”, Downloaded in 2016 from http://wwwhome.cs.utwente.nl/~hiemstra/papers/IRModelsTutorial-draft.pdf.

[12] W. B. Croft, “Knowledge-based and Statistical Approaches to Text Retrieval”, The IEEE Expert, Vol. 8, No. 2, pp. 8-12, 1993.

[13] Amit Singhal, "Modern Information Retrieval: A Brief Overview", Downloaded from in 2016 from http://www.gib.fi.upm.es/sites/default/files/irmodeling.pdf.

[14] Jelita Asian, “Effective Techniques For Indonesian Text Retrieval”, Ph.D. Thesis, the School of Computer Science and Information Technology, RMIT University, Melbourne, Victoria, March 2007.

[15] K. Tamsin Maxwell, “Term Selection in Information Retrieval”, Ph.D. Thesis, Institute for Communicating and Collaborative Systems, University of Edinburgh, January 2014.

[16] Lilac Al-Safadi, Mai Al- Badrani, and Meshaal Al- Junidey, “Developing Ontology for Arabic Blogs Retrieval”, International Journal of Computer Applications, Vol. 19, No. 4, pp. 41-46, April 2011.

[17] Jaffar Atwan, Mosnizah Mohd, and Ghassan Kanaan, “Enhanced Arabic Information Retrieval Light Stemming and Stop-Words”, M-CAIT2013, CCIS378, Springer Verlag Berlin Heidelbeg, pp. 219-228, 2013.

[18] Jaffar Atwan, Mosnizah Mohd, Hasan Rashadeh, and Ghassan Kanaan, “Semantically Enhanced Pseeudo Relvence Feedback For Arabic Information Retrieval”, Journal of Information Science, Vol. 42, No. 2, pp. 246-260, 2016.

[19] Mohamed Wedyan, Basim Alhadidi, and Adnan Alrabea, “The Effect of Using an Arabic Information Retrieval System”, International Journal of Computer Science Issues, Vol. 9, No. 6, pp. 431-435, November 2012.

[20] Waseem Alnomima, Ibrahim F. Moawad, Rania Elgohary and Mostafa Atef, “Ontology Based Query Expansion for Arabic Text Retrieval”, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 8, pp. 223-230, 2016.

[21] Cheng-Hui Huang, Jian Yin, and Dong Han, “An Improved Text Retrieval Algorithm Based on Suffix Tree Similarity Measure”, Springer-Verlag Heidelberg, ICICA, 2010, pp. 150-157, 2010.

[22] Susan Dumais, Edward Cutrell, J. J. Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins, “A System for Personal Information Retrieval and Reuse”, SIGIR, Toronto, Canada, 2003.

[23] Mohamed A. Abdelhadi, Tirveedula Gopi Krishna, and Ghassan Kanann, “A New Developed Model for Arabic Information Retrieval System Based on Knowledge Base System”, International Journal of Emerging Research in Management and Technology, Vol. 2, No. 11, pp. 1-7, November 2013.

[24] Roi Blanco Gonzalez, “Index Compression for Information Retrieval System”, Ph.D. Thesis, University of A Corunna, 2008.

[25] Kolikipogu Ramakrishna and B. Padmaja Rani, “Study of Indexing Techniques to Improve the Performance of Information Retrieval in Telguw Language”, The International Journal of Emerging Technology and Advanced Engineering, Vol. 3, No. 1, pp. 1-10, January 2013.

[26] Moon Soo Cha, So Yeon Kim, Jae Hee Ha, Min-June Lee, Young-June Choi, and Kyng Ah Sohn, “Topic Model Based Approach for Improved Indexing Content based on Document Retrieval”, International Journal of Networked and Distributed Computing, Vol. 4, No. 1, pp. 55-64, January 2016.

[27] Yang Wei, Jinmao Wei, Zhenglu Yang, and Yu Liu, “Joint Probability Consistent Relation Analysis For Document Representation”, Springer International Publishing Switzer Land, LCNS 9642, pp. 517-532, 2016.

[28] Stefan Pohl, “Boolean and Ranked Information Retrieval For Biomedical Systematic Reviewing”, Ph.D. Thesis, Department of Computer Science and Software Engineering, University of Melbourne, Victoria, Australia, Feb. 2012.

[29] Xiao Wei, Jun Zhang, Daniel Dajun Zeng, and Qing Li, “A Multi-Level Text Representation Model Within Background Knowledge Based on Human Cognitive Process For Big Data Analysis”, Cluster Computing, Vol. 19, pp. 1475-1487, 2016.
[30] Yang Wei, Jinmao Wei, and Hengpeng Xu, “Context Vector Model for Document Representation: A Computational Study”, Springer International Publishing Switzerland, NLPCC 2015, LNAI 9362, pp. 194-206, 2015.
[31] Leemon Baird, and Donald H. Kraft, “A New Approach for Boolean Query Processing in Text Information Retrieval”, Downloaded From the Internet in 2017 From the Website http://leemon.com/papers/2007bk.pdf.
[32] S. Ruban, S. Behin Sam, Lenita Veleza Serrao, and Harshitha, “A Study and Analysis of Information Retrieval Models”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, No. 7, pp.1-7, October 2015.
[33] Tareq Z. Ahram, “Information Retrieval Performance Enhancement Using the Average Standard Estimator and the Multi-Criteria Decision Waited Set of Performance Measures”, Ph.D. Thesis, Department of Industrial Engineering and Management Systems, College of Engineering and Computer Science, University of Central Florida Orlando, Florida, 2008.
[34] Simon Jonassen, and Svein Erik Bratsberg, “Improving the Performance of Pipelined Query Processing with Skipping—and its Comparison to Document Wise Partitioning”, Downloaded From the Internet in 2017 From the Website https://link.springer.com/chapter/10.1007/978-3-642-35063-4_1.
[35] Balwinder Saini, Vikram Singh, and Satish Kumar, “Information Retrieval Models and Searching Methodologies: Survey”, Downloaded in 2016 From the Website https://www.researchgate.net/publication/274837522.
[36] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, “An Introduction to Information Retrieval”, Cambridge University Press, 2009.
[37] Michael McCandless, Erik Hatcher, and Otis Gospodnetic, "Lucien in Action", The 2nd Edition, 2005, Downloaded From the Internet in 2017 From the Website http://www.apache.org/licenses/LICENSE-2.0

BIGRAPHY:

Prof. Dr. Nawal El-Fishawy received the Ph.D degree in mobile communications, Faculty of Electronic Eng., Menoufia University, Menouf, Egypt, in collaboration with Southampton University in 1991. Now she is the head of Computer Science and Engineering Dept., Faculty of Electronic Eng. Her re-search interest includes computer communication networks with emphasis on protocol design, traffic modeling and performance evaluation of broadband networks and multiple access control protocols for wireless communications systems and networks. Now she directed her research interests to the developments of security over wireless communications networks (mobile communications, WLAN, Bluetooth), VOIP, and encryption algorithms. She has served as a reviewer for many national and international journals and conferences.

Prof. Dr. Mohamed Nour Elsayed is a professor of computer engineering at the Electronics Research Institute, Cairo. He was graduated from the Computer Department at the Faculty of Engineering, Ain Shams University in 1980. He obtained his M.Sc. and Ph.D. in 1987 and 1993 respectively. He taught more than twenty-years ago at the American University in Cairo (AUC) as a part-time instructor. He taught also five years ago at Princess Nourah University, Riyadh, KSA. He was the head of the Department of Research Informatics as well as the Vice-President of the Electronics Research Institute, Cairo. He is an IEEE member and a reviewer of some national and international computer journals. The areas of his interest include; but not limited to; high performance computing, computational linguistics, and artificial intelligence applications.

Dr. Gamal M. Attiya graduated in 1993 and obtained his M.Sc. degree in computer science and engineering from Menoufia University, Egypt, in 1999. He received PhD degree in computer engineering from University of Marne-La-Vallée, Paris-France, in 2004. He is currently associate professor at Computer Science and Engineering department, Faculty of Electronic Engineering, Menoufia University, Egypt. His main research interests include distributed computing, allocation and scheduling, cloud computing, Big Data analysis, computer networks and protocols.

Dr. Maha Saad Tolba is a lecturer of Computer Engineering at the Faculty of Electronic Engineering, Menofia University. She was graduated from the Department of Computer Science and Engineering, Faculty of Electronic Engineering, Menoufia University in 1997. She obtained her M.Sc. and Ph.D. in 2006 and 2011 respectively. The areas of her interest include; but not limited to; computer networks, information security, and information technology.

Eng. Ayat Elnahaas is a research assistant at the Department of Research Informatics, Electronics Research Institute, Cairo. She was graduated from the Faculty of Electronic Engineering, Menofia University in 2013. Currently; she is working in her M.Sc. in the area of Arabic text processing. The research areas of her interest are: computational linguistics, and information technology.

تحليل آداء وتعديل نموذج استرجاع المعلومات العربية إعتمادا على تمديد استفسار المستخدم

آيات النحاس*[1]، نوال الفيشاوى**[2]، محمد نور*[3] ، جمال عطية**[4] ، مها طلبة**[5]
قسم بحوث المعلوماتية، معهد بحوث الإلكترونيات
القاهرة- جمهورية مصر العربية
[1]eng_ayatelnahas@yahoo.com; [3]mnour@eri.sci.eg
***قسم هندسة وعلوم الحاسب- كلية الهندسة الإلكترونية*
منوف- جامعة المنوفية- جمهورية مصر العربية
[2]nelfishawy@hotmail.com, [4]gamal.attiya@yahoo.com; [5]maha_saad_tolba@yahoo.com

الملخص:
تهدف عملية استرجاع المعلومات إلى إيجاد الوثائق والنصوص المناسبة والتى تلبى رغبات استفسار المستخدم. يهدف هذا العمل البحثى إلى تحليل وتدقيق أحد النماذج لاسترجاع المعلومات، ومن ثم سيتم عرض العناصر الأساسية لذلك النموذج مثل تجميع الوثائق، تمثيل استفسار المستخدم، عمل الفهرسة، وكذا عمل المضاهاة. ومن منطلق أن اللغة العربية هى أحد اللغات الهامة فى اللغات الطبيعية التى يتعامل بها العالم، فإنه قد تم استخدام مجموعة من الوثائق العربية لاختبار آداء ذلك النموذج االذى تم إختياره. وعلى ذلك فإن هناك بعض العمليات سيتم إجراؤها لتسهيل عملية المضاهاة بين الكلمات الدالة لاستفسار المستخدم مع العناصر التى تصف كل وثيقة أو نص عربى، ومن أمثلة تلك العمليات: عملية تجزئة النص العربى إلى كلمات Tokens، استبعاد الكلمات التى لاتؤثر فى عملية استرجاع النصوص العربية Removal of Stopwords، وكذا إيجاد أصل الكلمة العربية Stemming بعد تجريدها من أحرف الزيادة سواء القبلية أو البعدية.
سيقدم هذا العمل أيضا إقتراحين لتعديل النموذج المستخدم رغبة فى تعزيز كفاءتة وتحسين آدائة. هذا ويعتمد التعديل على تمديد استفسار المستخدم. فالمقترح الأول يقوم بتمديد الكلمات الدالة لاستفسار المستخدم وذلك من خلال إضافة كلمات جديدة تعتمد على المعانى الدلالية للكلمات الأصلية لاستفسار المستخدم عن طريق الاستعانة بالقاموس العربى للحصول على معانى تلك الكلمات Semantics. ويقوم المقترح الثانى بتمديد الكلمات الدالة فى استفسار المستخدم بإضافة بعض الكلمات الدالة المصاحبة لبعض النصوص المسترجعة التى يراها المستخدم متوافقة مع استفسارة الأصلى، وأيضا يقوم هذا المقترح باستبعاد أى كلمات مصاحبة لبعض النصوص المسترجعة التى لايرحب بها المستخدم، وهذا ما يطلق عليه Relevance Feedback.
إضافة لما تقدم فإن نموذج استرجاع المعلومات سيتم تطبيقة واختباره وتقييمة قبل وبعد التعديلين المشار إليهما سابقا. هذا وسيتم تقييم آداءالنموذج والتعديلات التى ستجرى عليه من خلال عدد من المعايير مثل معيارالدقة، إعادة الاسترجاع، ومقياس F أو مايعرف باسم Precision, Recall, and F-measure.
هذا وتشير نتائج التجارب التى تم إجراؤها إلى أهمية عملية الحصول أصل الكلمات العربية، استبعاد الكلمات غير المؤثرة، وكذا أهمية نموذج الفراغ المتجهى. ويعتبر آداء النموذج باستخدام التعديل الأول أفضل من آداء النموذج الأصلى الذى تبنتة الدراسة بحوالى 27%، بينما وصلت نسبة التحسن إلى مايقارب 14% باستخدام التعديل الثانى مقارنة أيضا بالنموذج الأصلى. ومما تجدر الإشارة به هنا هو أن نسبة التحسن كانت أفضل بضم التعديلين سويا والتى وصلت فى حدود 15% إلى 35% مقارنة بالنموذج الأصلى.

الكلمات الدالة: الوثائق العربية، آعمال الفهرسة، نموذج المتجة الفراغى، تمديد استفسار المستخدم، المعانى الدلالية للكلمات الدالة، التغذية الخلفية المناسبة.