International Journal of Information Retrieval Research
Volume 12 • Issue 1

Hybrid Query Expansion Model Based on Pseudo Relevance Feedback and Semantic Tree for Arabic IR

Ahmed Cherif Mazari, Computer Science Department, University of Biskra, Algeria
Abdelhamid Djeffal, Computer Science Department, University of Biskra, Algeria

ABSTRACT

In this paper, the authors propose and readapt a new concept-based approach of query expansion in the context of Arabic information retrieval. The purpose is to represent the query by a set of weighted concepts in order to identify better the user's information need. Firstly, concepts are extracted from the initially retrieved documents by the pseudo-relevance feedback method, and then they are integrated into a semantic weighted tree in order to detect more information contained in the related concepts connected by semantic relations to the primary concepts. The authors use the "Arabic WordNet" as a resource to extract, disambiguate concepts, and build the semantic tree. Experimental results demonstrate that measure of MAP (mean average precision) is about 10% of improvement using the open source Lucene as IR System on a collection formed from the Arabic BBC News.

KEYWORDS

Arabic Information Retrieval, Arabic Word Net, Concept-Based Query Expansion, Pseudo Relevance Feedback, Semantic Tree

INTRODUCTION

The classical Information Retrieval Systems (IRSs) mainly use words to represent the content of documents and queries, hence, the matching is carried out by using these words on lexical level rather than on semantic level, in other words, when the IR system takes a query, it simply retrieves documents that contain the query words without considering the semantics behind them. Furthermore, many users give different words to refer to the same concept or they only provide few words and do not explicitly describe the information need. Therefore, Users' query words may be quite different and not specific enough to the ones used in the documents for describing the same semantics. The user may then get many of irrelevant documents in the result set. Clearly, such vocabulary gaps make the retrieval effectiveness non-optimal and decrease the IRS performance.

Query expansion (QE) is one of the proposed approaches used to solve problems mentioned above, based on the following principle; more the number of keywords in the query is greater, more the information need is well described, since it certainly includes a greater number of index keywords that represent relevant documents (Xu & Croft, 2000). Indeed, it consists of adding new words or terms into the original query in order to improve the retrieval performance.

In the past two decades, several query expansion techniques were developed, i.e. the strategy of the *direct reformulation,* it consists by adding new terms to the initial query that are extracted from external resources such as, ontologies using concepts (Bhogal, Macfarlane, & Smith, 2007) or thesauri

DOI: 10.4018/IJIRR.289949

as Wikipedia (Li, Luk, Ho, & Chung, 2007). In the strategy of *indirect reformulation*, this consists by adding new terms extracted from a list of documents already considered relevant, called Relevance Feedback when it is supervised and documents are chosen by the user, or Pseudo Relevance Feedback (PRF) when documents are automatically selected by the system (Xu & Croft, 1996).

Although the traditional methods for query expansion have got many improvements, however, they have to be used carefully, because, it might degrade the performance of the IRS, This is due to the incorrect choice of terms in which implies the divergence to user's information needs (Cronen-Townsend, Zhou, & Croft, 2004). In addition, existing approaches mainly focus on isolated keywords without regard for the query's semantics to expand it integrally, i.e. they do not exploit semantic relations between keywords. They also consist by adding directly terms from the external resource, for which results are bound to be less subjective.

This paper presents a new hybrid approach of query expansion (QE) for Arabic Information Retrieval (IR); in which the authors demonstrate how the PRF expansion can be combined to an external resource such as the Arabic WordNet (AWN) to enhance QE process, in selection and weighting of expansion terms. Thus, the combined approach aims to represent the user's information need by a semantic tree whereby each node is an expansion concept including one or several candidate terms. In other words, is to build a weighted concept hierarchy following the user's information need and semantic relations of keywords.

The proposed approach is subdivided into three major stages. First, the query is pre-processed, which usually includes normalization, elimination of stopwords, and stemming, etc. Whereby, simple and compound terms are extract. These terms are then matched on the AWN to extract and disambiguate appropriate concepts. The concept list is also given to the IR system to return a set of documents ranked according to relevance criterion of the system. Thereby, the most important terms are extracted from the top-R documents of the Pseudo Relevance Feedback (PRF) process. Second, the query expansion procedure starts by building a set of initial semantic trees, in which roots are the original keywords and their synonyms found in the AWN resource, all these trees are integrated and consolidated to one big tree; by adding relations to concepts that are semantically related, and deletion of concepts that are semantically distant.

Finally, a weight is assigned to each node of the integrated tree. These weights allow selecting candidate concepts according a given threshold in order to generate the best candidate terms for the QE process.

The paper is structured in four sections; related work is provided in section 2 in order to identify techniques and methods that motivated us to overcome the weakness and increase the performance of Arabic Information Retrieval. Section 3 presents the proposed approach for query expansion. Results of experiments are presented in section 4. The last section brings the main conclusion.

RELATED WORK

This work is related to three research fields that are being actively furthered in the discipline of Information Retrieval; the Query Expansion techniques using the PRF method combined to external resource (such as WordNet), the extraction, disambiguation and weighting of concepts, and adapting of the Arabic language to IR.

Query Expansion Approaches

Query expansion is one of the important techniques to overcome the word mismatch problems and to be closer to user's need in information retrieval. It is categorized into three main categories based on the methods and resources of expansion: First, the approaches based on the information derived from a set of documents initially retrieved (Local approaches or feedback). Second, the approaches based on global information derived from all the documents in the collection (Global approaches). Third, the approaches based on knowledge structures like corpora, thesauri, dictionaries or the combination

of them. Several approaches to query expansion have been studied in the past. The earliest trials were carried by (Jones, 1971) who used clustered words based on co-occurrence in documents to expand the query. Then, global and local analysis techniques were used by (Xu & Croft, 2000). In this regard, an interesting survey presented in (Carpineto & Romano, 2012).

Global Analysis

The global analysis (GA) of corpus is carried out to get the semantic correlation between all words of the corpus and original keywords of the query, and then words that have more similarities than a given threshold are selected as candidates for query expansion. Experimental results on the TREC collection showed the effectiveness of the GA approach proposed by (Hu, Deng, & Guo, 2006).

Local Analysis

The Local Analysis (LA) is carried only in the top retrieved documents for the original query known as relevant feedback documents, the candidate words of expansion are selected by computing similarities from analysing on text passages (a text window with a fixed size) as proposed in (Xu & Croft, 1996). LA can usually get higher performance than GA, but the result of LA is determined by the initial retrieval (Huang, Wang, & Zhang, 2011). Deep investigations on feedback methods have shown better results, such as, the work of Colace et al. (2015), where the authors proposed a query expansion method that automatically extracts a set of weighted word pairs from a set of topic-related documents provided by the relevance feedback and basing on probabilities. The performed evaluations demonstrated the effectiveness with respect to the baseline. Also, the work presented by Karisani et al. (2016) has proposed a method to identify and re-weight informative query terms, by examining the similarity of top documents and weighting them based on their context. The analysis of results obtained indicate that the suggested method is capable of identifying the most important keywords even in short queries and it improves retrieval performance around of 7% of MAP over traditional query term re-weighting methods.

Thesaurus

The thesauri, as external resources, have been widely used in the query applications. the work of Stairmand (1997) based on the WordNet for query expansion concluded that the improvement was limited by the coverage of the WordNet, thus the expansion process is closely related to the richness of this resource. For works based on WIKIPEDIA, Li et al. (2007) developed a method of query expansion based on this resource by using categories of documents, in which, the weight of the category is assigned to initial query keywords. Khoury (2011) used the same resource for identifying the synonymous expressions and linguistic entities that are semantically similar to original keywords and adding them to query set before starting the retrieval process. Aggarwal and Buitelaar (2012) are also extracted knowledge from WIKIPEDIA and DBPEDIA for query expansion. Egozi et al. (2011) discussed the query expansion method that combines both thesaurus and local method, by building a local thesaurus from the returned documents of the initial search and use it to expand the query.

Concept-Based Methods

Several works in the literature are devoted to the study of concept-based query expansion, which have been extracted form ontologies. Thus, when the researchers use concepts rather than single terms they try to express implicit senses that are not given by isolated terms, therefore a concept is represented as a set of adjacent terms. Liu et al. (2009) proposed a sense recognition of hyponyms based on concept space. They used the contexts of hyponyms and the weights of feature words to construct a hyponyms-word vector space. Aseervatham (2009) also studied a Concept Vector Space Model (CVSM), which uses linguistic prior knowledge to capture the meanings of the documents. In addition, the study of Huang et al. (2011) presented a new approach to QE, in which the idea is to construct a “Tree of Associational Semantics Model” TASM, and select candidate keywords from

the tree. Their experiments show that the result of this approach is better than the traditional method based on tf*idf. Recently, Aklouche et al. (2019) presented a PRF method based on co-occurrence graphs by measuring term-term similarity.

Related Works to Arabic

Arabic is the language that has many properties; First, Arabic language consists of 28 letters, 16 of them have one dot, two or three dots. Second, the Arabic letters' shape changes depending on their position in the word (in the beginning, the middle or the end). Third, Writing and reading are from right to left. Fourth, diacritics (Tashkil) are optional: written Arabic text can be fully diacritized (e.g., Classical Arabic including historical and liturgical texts), partially diacritized, or entirely undiacritized. In this study, the authors interest in Modern Standard Arabic (MSA). MSA is the official language of the Arab World. MSA is the primary language of the media and education. MSA is syntactically, morphologically and phonologically based on Classical Arabic (Habash, 2010). Fifth, grammatical flexibility; words may be arranged in many different ways (Khatib, 1997). The written language has no capital letters for proper names, as names of people, geographical positions, cities, countries, months, and days of the week, etc. which creates increased ambiguity and especially complicates such tasks as text processing, automatic indexing, information extraction, named entity recognition. Arabic WordNet (AWN) is one of the resources that have been developed in order to overcome these characteristics. It is based on the Princeton WordNet (PWN), which includes 23481 words and 11269 synsets (Elkateb et al., 2006). The AWN is often used in the concept-based IR, which allows the retrieval of documents based on meaning expressed by the terms of the query. In this approach, the authors use it by applying different techniques.

Several studies investigating query expansion have been carried out on Arabic IR. Shaalan et al. (2012) suggested a method for query expansion based on similarity of terms to improve Arabic IR using Expectation-Maximization. Their experiments have shown that QE considers the similarity of terms enhances precision and retrieves documents that are more relevant. The work of Mallat et al. (2013) proposed an enrichment of Arabic queries based on linguistic and contextual to improve the performance of the information retrieval systems. The goal of the method is to generate a descriptive list containing a set of linguistic lexicons to assign to each significant term in the query, and a second enrichment consists in integrating contextual information derived from the corpus. Mahgoub et al. (2014) introduced a query expansion approach using an ontology built from Wikipedia pages to improve search accuracy for Arabic language.

The Arabic WordNet (AWN) has been also applied as an external resource to improve information retrieval recently. Mazari et al. (2013) used this resource for conceptual Arabic document indexing. Abderrahim (2014) has examined and compared two techniques for query expansion in Arabic IR; concept-based QE using Arabic WordNet and Pseudo Relevance Feedback (PRF), the results obtained shown that the PRF is better than concept-based retrieving and PRF reformulation can improve the performance of an Arabic IRS about 4%. Abbache et al. (2018) described an automatic query expansion (AQE) method using the Arabic WordNet and Association Rules within the context of Arabic Language. Based on the assumption that words that tend to occur together in documents are likely to have similar meanings and hence can be associated, the association can occur between query's terms and the synonym. On the other hand, for indexing, the work of El Mahdaouy et al. (2018) has demonstrated that Word embedding-based language models significantly outperform the semantic indexing approach based on Arabic WordNet. This is explained by the fact that the Arabic WordNet is limited for covering the Arabic language and when it was used alone as an external resource. Several scenarios of query reformulation and document indexing of Arabic IR are evaluated based on morpho-semantic knowledge graph in (Bounhas, Soudani, & Slimani, 2020).

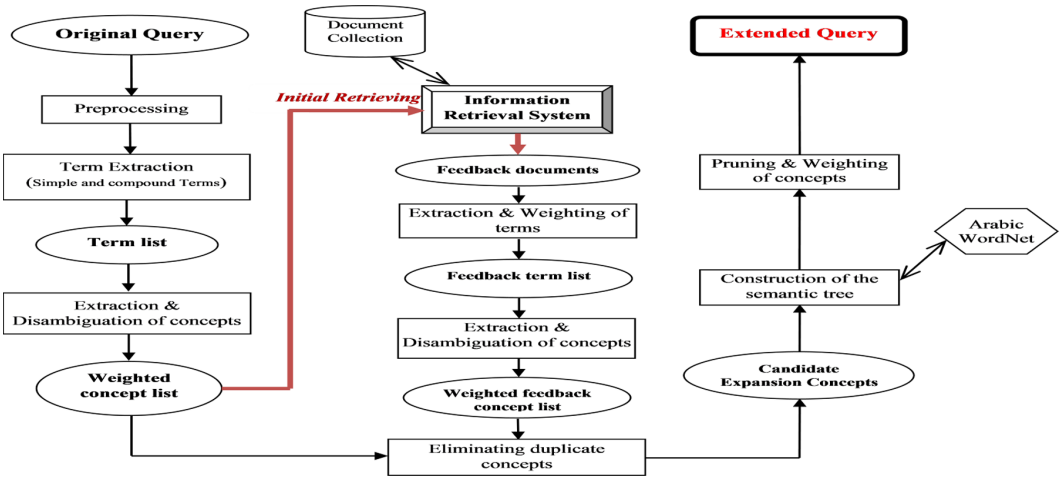
However, this proposed approach is a new work against the previous studies; it combines the knowledge from Arabic WordNet with the pseudo relevance feedback technique to represent the user's information need by a semantic tree.

PROPOSED APPROACH

In this section, the authors describe components that they used to study and implement their proposed approach for query expansion in the context of Arabic IR. It consists of a combination of a local approach and an approach based on an external resource. In this respect, the authors propose to combine the Pseudo Relevance Feedback (PRF) method and the Arabic WordNet resource to represent the user's need by a semantic tree. The process must be completely independent of the information retrieval system (IRS), in other words, the expansion process does not modify the internal representation of the documents or the index. The main contribution of the process is then -by formulating a richer and more precise query-.

The processing flow of the proposed approach illustrated in the Figure 1. is described in the following main steps; First, preprocessing and extraction of concepts from the original query, Second, initial retrieving by the PRF method, Third, the building of the semantic tree by using the Arabic WordNet (AWN) as a semantic resource, Finally, selection of candidate expansion keywords.

Figure 1. Flow chart of the proposed approach



Pre-Processing of the Original Query

Arabic is a language with a high degree of inflection. The preprocessing of the user's initial query represents a big challenge due to the features of Arabic linguistics and morphology. It consists mainly of:

Tokenization

Queries are in natural language form. In this phase, the stream of the query text is broken up into words, symbols, or other meaningful elements called tokens, basing on spaces and on the delimiters represented by the punctuation and the non-alphabetic characters.

Normalization

By removing punctuation, some diacritics, special characters, foreign letters, white spaces and numbers, and replacing إ, أ, and آ by ا, ي by ى, and ة by ه.

Eliminating Stopwords

Arabic stopwords are available in large numbers, due to richness of lexical tokens; they include some of the grammatical links such as the definite article "the" (ال), attached and separated prepositions, conjunctions, interrogative words, negative words, exclamations, calling letters, adverbs of time and place. They also include all the pronouns, the demonstratives, the subject/object pronouns, the five distinctive nouns, some numbers, the additions and the verbs. In this work, the authors use the general stopword list created and tested by (El-khair, 2006). It was collected based on Arabic language characteristics; this general stopword list has a large number of empty words (consisting of 1377 words). The reason for choosing this list is due to the best performance for Arabic Information retrieving using the BM25 (Robertson & Zaragoza, 2009).

Stemming

Stems are often used to improve retrieval performance because they reduce variants of the same root word to a common concept, whence the automatic stemming is getting the root of the word by deleting the prefixes and/or the suffixes. In this work, the Khoja Stemmer is applied, this is the one of the tools widely used in Arabic language (Khoja & Garside, 1999).

Extraction of Terms

There are two types of terms: simple terms composed of a single word and compound terms composed of a series of words.

Simple Term Extraction

A simple term is considered as a non-stopword that could be used to describe the content in a query, so the simple term set is the list of terms of the preprocessed query.

Compound Term Extraction

It allows reducing the ambiguity while assigning a term to a concept in the AWN. Sometimes, query terms contribute together to form a unique concept even if they refer to different senses when they are taken separately. For that, firstly, the entire query is considered as one term. Then, the technique based on n-gram process where n varies from (size (query) - 1) descending to one, in order to detect all compound terms that contain the query.

For example: to disambiguate the compound term "*Bethlehem*" "مح لم تي بي", which is composed of two simple terms q_1=بيت and q_2=لحم . There are seven senses for q_1 and one sense for q_2, so a total of 7*1 possible combinations. Hence, the AWN is used to assign the appropriate sense to the extracted compound term.

Extraction and Disambiguation of Concepts

Concept Identification

In this step, the resulted list of simple and compound terms from the previous step is projected onto AWN. If an unambiguous term (belongs to a unique synset of the AWN) is obtained, then its related concept is extracted and added to the concept list. Else, if an ambiguous term (belongs to more than one synset) is obtained, then it must be disambiguated to have the right concept of the given term.

Disambiguation Procedure

In order to disambiguate an ambiguous concept (represented by a synset), the semantic similarity between concepts and the query must be calculated. Several methods are used by authors in related work like the semantic distance between concepts of the ontology and their weight scores such as works presented by (Baziz, Boughanem, & Traboulsi, 2005; Boubekeur, Boughanem, & Tamine-Lechani, 2007). In this proposed approach, the authors propose to use the Jaccard coefficient formula (1), such

as the work of (Pal, Mitra, & Datta, 2014). Whereby, a score is assigned to each candidate concept as proposed in formula (2). The idea is to calculate the score between each ambiguous concept and the query, thus, the best concept that represents the term is the one that maximizes this score. This method is more important than if the query terms are independent.

$$Sim_{Jaccard}\left(t,q_i\right)=\left(\frac{r_{t,q_i}}{r_t+r_{q_i}-r_{t,q_i}}\right) \tag{1}$$

The Jaccard formula computes the similarity between two terms. Where: i is the query term, t is a term, r_i and $r_{i,j}$ denote, respectively, the number of documents in which term i occurs, and the number of documents in which i and j co-occur.

$$Score\left(C\right)=1/k\left(NC+\sum_{i=1}^{k}Sim_{Jaccard}\left(t,q_i\right)\right) \tag{2}$$

Where: *K* denotes the number of terms of the synset C, *NC* is the number of common terms between synset C and the initial query, *t* denote the term to disambiguate, *i* is a term in the synset C.

Concept Weighting

It is obvious that the concepts in the query should not have an equal weight i.e. terms are not of equal importance. Therefore, different weights must be assigned to various concepts depending on their importance in the query. Several methods are used for weighting. In this approach, the authors assign the previous weight scores to the given concepts, calculated by the formula (2).

Initial Retrieving by the PRF Method

The first expansion process use the Pseudo Relevance Feedback method, it consists in retrieving a collection of documents using the IRS requested by previously weighted concept terms.

Extraction and Weighting of Terms.

The top-R retrieved documents are assumed as relevant, they are selected as a corpus for extracting the set of new candidate terms. In this proposed approach, the authors suggest extracting the most representative terms. Hence, they calculate their proportions of frequencies given by the formula (3), and the ones that exceed a threshold (m) are selected as candidate terms.

$$p\_freq_i=\frac{freq_i}{Max\_freq} \tag{3}$$

Where: $freq_i$ denotes the frequency of the term *i* in the top-R retrieved documents and Max_freq denotes the higher frequency of terms in the top-R retrieved documents.

Extraction and Disambiguation of Concepts

In order to extract the PRF concepts, the authors consequently use the previously extracted terms that have the proportional frequencies exceeding the threshold (m). Whereby, the same concept identification and the same disambiguation procedure presented in section (*Extraction and disambiguation of concepts*) are applied.

Weighting of Feedback Concepts

For weighting these new concepts, the score calculated by the formula (2) is also assigned to each feedback concept.

Eliminating of Duplicate Concepts

A new list of weighted concepts is obtained, by adding these new weighted concepts from the PRF process to the previous concept list that means to have a duplicate concept possibility, which must be eliminated. Hence, the appropriate list of weighted concepts is elected.

Building the Semantic Tree

Based on the extracted candidate concepts Ci of the previous steps, a semantic tree is built with the aim for capturing the semantics of the user's need, in which is used to expand the corresponding query.

Initialization

First, a group of initial semantic trees are generated by using the AWN resource. Where, the initial candidate concepts C_i and their synonyms in roots are located. Second, for not diverging from the initial concept meanings, new concepts are then extended by using the hyponymy relation at two levels for the original concepts and at one level for their synonyms. All nodes, except the initial concepts, are retrieved from the AWN. The extension sets are generated according the following:

- $S_i = \{s_{ij} \mid 1£j£m\}$ denotes synonym set of c_i.
- $A_i = \{a_{iu} \mid 1£u£t\}$ denotes hyponym set of c_i.
- $B_{iu} = \{b_{iuv} \mid 1£v£p\}$ denotes hyponym set of a_{iu}.
- $D_{ij} = \{d_{ijf} \mid 1£f£z\}$ denotes hyponym set of $s_{ij.}$

Integrating

In another hand, using also the AWN, these initial trees are extended in departing upward based on the hyperonymy relation until finding a common node, which is represent the root of integrated global semantic tree, as shown in Figure 2.

In the following Figure 3, the authors show how the tree of the query *Q116* has been generated.

```
<num>116</num>
  <title>ةكرح ا تلمرت في ا دوسلا ن</title>
  <title_e>the rebel movement in Sudan</title_e>
```

Weighting and Pruning

To compute the average similarity between each node of the semantic tree and the original concepts, the authors propose formula (4) for weighting (average similarity) of a given node and formula (5) to calculate the semantic similarity between a concept node and an original concept based on the global analysis.

$$weight(n) = \frac{1}{k}\sum_{i=1}^{k} sim(n, c_i) \tag{4}$$

Where: *n* denote a node in the integrated tree except the initial concepts c_i;

Figure 2. Semantic tree

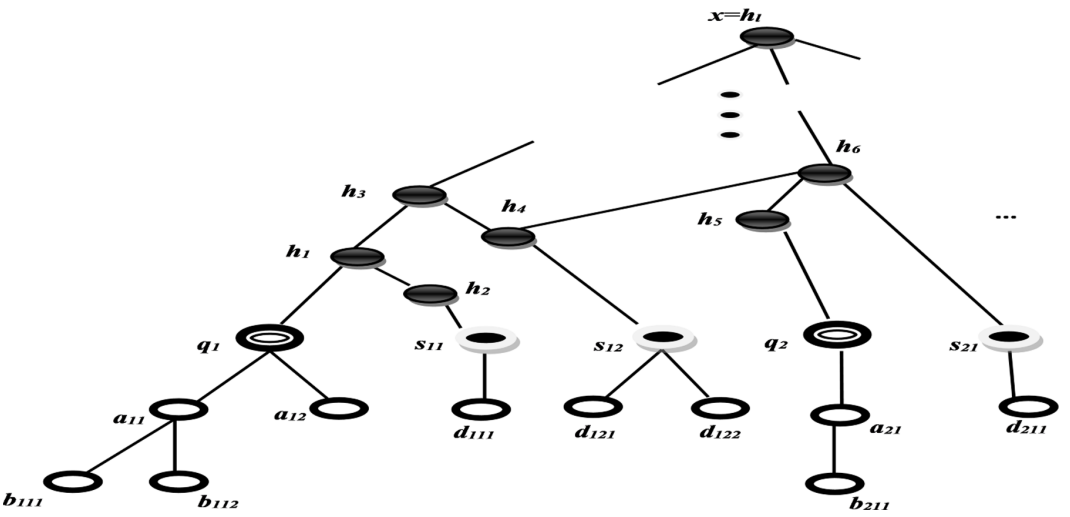
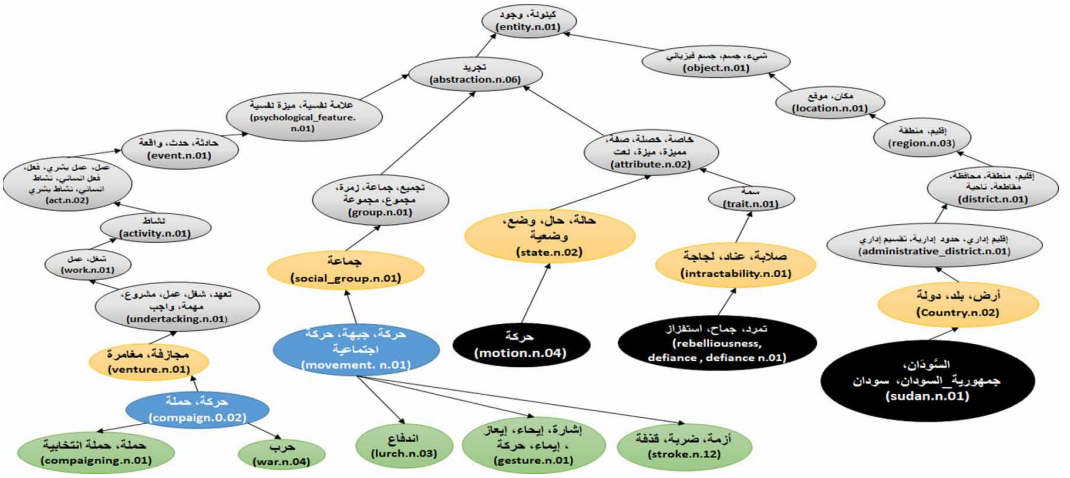


Figure 3. Semantic tree of the query Q116



$$sim(c_1, c_2) = \frac{common(c_1, c_2) + 0.5}{|freq(c_1) - freq(c_2)| + 0.5} \quad (5)$$

Where: common (c_1, c_2) is the number of common terms between c_1 and c_2, and freq(c) as formula (6).

$$freq(c) = \frac{1}{length(c)} \sum_{i=1}^{length(c)} count(i) \quad (6)$$

With: count(i) is the frequency of the term *i* in all documents of the collection.

Let *g* denote the number of nodes on the integrated tree except the original ones.

T is defined as a scheduled threshold (determined experimentally= 0.5), then the integrated tree is pruned and weighted as describe the following algorithm:

```
for l = 1 to g do
for i = 1 to k do
if  average similarity (n) ³ T) then
assign average similarity (n) to n as its weight
else
delete n from the tree;
end i;
end l.
```

Query Expansion

As result, the query reformulation process is defined by Q'. in which Q' = Q È Q1 È Q2, where Q represents the initial query, Q1 is the set of keywords corresponding to concepts which are extracted by PRF and Q2 is the set of keywords of concepts that are defined by the semantic tree. A corresponding weight *w* is assigned to each keyword *q* of the previous sets, before sending the expanded query to the IR System, as follow:

if q is an original keyword, i.e., it is from Q, then weight w = 1, because the original keywords have the best indication of user's interest;

if q is a keyword corresponding to concepts which are extracted by PRF, the weight w= PRF's weight multiplied by 0.5.

if q is a keyword of a concept from the semantic tree, the weight is calculated by multiplying their node's weights and 0.5.

EXPERIMENT AND DISCUSSION

In order to assess the effectiveness of the approach presented above, experiments are carried out according to the TREC protocol, using the precisions at p@5, p@10, p@15, P@20, p@100 and MAP (Mean Average Precision). Experimental design and results are described in following.

Experimental Design

The experiments are conducted on the corpus of journalistic texts consisting of 4763 articles recovered from the Arabic BBC News (Arabic-Corpora, 2010), created by (Saad & Ashour, 2010). This corpus covers various areas as shown in Table.1. It contains 1,860,786 words and 106,733 keywords after removing stopwords on seven topics.

The evaluation is performed on a set of 43 queries of various topics hosted by (list-of-queries, 2019), which the relevance judgments are known. Figure 4. shows a sample of some queries.

Furthermore, the optimal parameters values through the experiment are as follow, in section (*Extraction and disambiguation of concepts*) for initial retrieving by the PRF method; the Top-R=10 documents and the threshold m=0.6.

Experimental Results

In order to study and analyse the results; the authors have segmented the experimentation steps into three search tests by using Lucene as IR System (Lucene, 2019). Starting with the baseline run (43 original queries), then two query expansion methods are tested; concept-based Query Reformulation (by using weighted and extracted concepts from AWN resource and add them to the initial query),

Table 1. Corpus

Topic		# of text Doc
Middle East News	طسوألا قرشلا راخبأ	2356
World News	ملاعلا راخبأ	1489
Economy and Business	ادقتصاو أعمال	296
Sport	ةضايرلا	219
International Press	ةيملاعلا فاحصلا	49
Science & Technology	عولم وتكنولوجيا	232
Arts & Culture	فنون وثقافة	122

Figure 4. Sample of queries

```
  </query>
- <query>
    <num>119</num>
    <title>التدخلات السياسية لباراك أوباما في العالم </title>
    <title_e> Barack Obama's political interventions in the world </title_e>
  - <keywords>
      <keyword>باراك أوباما</keyword>
      <keyword_e> Barack Obama </keyword_e>
      <keyword> سياسة أوباما </keyword>
      <keyword_e> Obama's politic </keyword_e>
      <keyword> العالم</keyword>
      <keyword_e> world </keyword_e>
    </keywords>
    <relev_doc>106</relev_doc>
  </query>
- <query>
    <num>120</num>
    <title>إنسحاب القوات الأمريكية من العراق </title>
    <title_e> the withdrawal of US forces from Iraq </title_e>
  - <keywords>
      <keyword>إنسحاب </keyword>
      <keyword_e> withdrawal </keyword_e>
      <keyword> القوات الأمريكية </keyword>
      <keyword_e> US forces </keyword_e>
      <keyword>العراق</keyword>
      <keyword_e> Iraq </keyword_e>
    </keywords>
    <relev_doc>35</relev_doc>
  </query>
- <query>
    <num>121</num>
    <title>سياسة دبي لتجاوز الأزمة الاقتصادية </title>
    <title_e> Dubai policy to overcome the economic crisis </title_e>
  - <keywords>
```

and by applying techniques of the proposed approach. Thus, the authors calculate the precision of these queries at various points p@i and MAP of documents retrieved by Lucene.

The results obtained are shown in Table.2, and depicted in Figure 5.

Results

For better understanding the effectiveness of the approach, these results are described by percentage comparisons (%), into two categories; retrieving by concept-based query reformulation compared to baseline retrieving (without reformulation) and retrieving by query reformulation based on the semantic tree compared to baseline retrieving (without reformulation), as shown in Table.3.

According to these values, the work concludes the following: a poor performance when only using the query expansion by weighted concepts, which are -15.93%, -10.71%, -14.19%, -9.05%, -7.92% and -6.45% for p@5, p@10, p@15, P@20, p@100 and MAP respectively. This is mainly due to the coverage of the AWN to Arabic language, in which only 68% of the query keywords are found in the AWN and are conceptualized.

In the second test, there is a stable improvement as the QE by this approach compared to baseline, which are 7.61%, 5.86%, 7.20%, 7.73%, 7.55% and 10.08% for p@5, p@10, p@15, P@20, p@100

Table 2. Experimental result values

Precision at @x documents	Without Reformulation	Query Reformulation by Weighted Concepts	Query Reformulation by the Approach
P@5	0.565	0.475	0.608
P@10	0.495	0.442	0.524
P@15	0.472	0.405	0.506
P@20	0.453	0.412	0.488
P@100	0.265	0.244	0.285
MAP	0.248	0.232	0.273

Figure 5. Results of different search methods

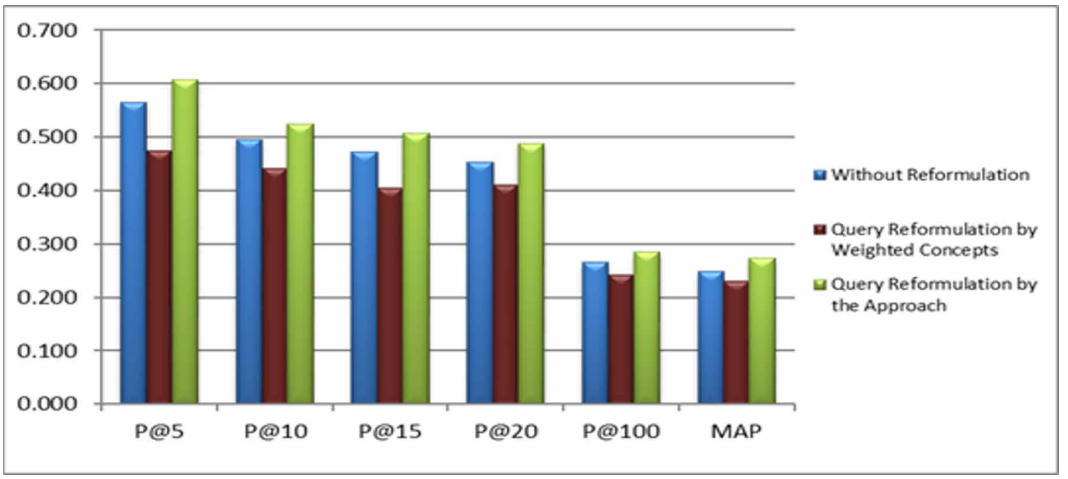


Table 3. Comparison of different search methods

	P@5	P@10	P@15	P@20	P@100	MAP
QR by weighted concepts % Baseline	-15.93%	-10.71%	-14.19%	-9.05%	-7.92%	-6.45%
QR by the Approach % Baseline	7.61%	5.86%	7.20%	7.73%	7.55%	10.08%

and MAP respectively. The experimentation permitted to show that the query reformulation using a semantic tree built following this proposed approach converge more to describe user information need.

Discussion

Following the previously described experimental process demonstrates improved performance of MAP around of 10% when using the semantic tree to capture the user's need, tested on 43 queries against: First, the concept-based method using the Arabic WordNet and Pseudo Relevance Feedback presented by (Abderrahim, 2014). In which it slightly improves the performance of the IRS about

4%. Conversely, the method using the AWN and Association Rules in Arabic IR based on relations to select the appropriate synonyms of terms described by researchers in (Abbache et al., 2018). They showed an increasing of performance around 13% of MAP, but they only tested on set of short keyword queries and sub-corpora. Second, Karisani et al., (2016) proposed a method, in a local analysis, for identifying and re-weighting informative query terms, that improves retrieval performance around of 7% of MAP over traditional query term re-weighting methods.

This proposed work give a solution to represent the keywords as a semantic hierarchy of the query, thus the user's need. Moreover, it also shows difficulties and limitations when the AWN resource is blindly used to expand queries.

CONCLUSION

The query expansion technique can often bridge vocabulary gaps between queries and documents, which can improve the performance of retrieving relevant documents. In this context, this study propose a new approach to query expansion for Arabic IR, through a semantic tree representation. This tree is built from the query keywords and their corresponding extensions of child and parent concepts, for this purpose, the AWN resource is used at the same time for the concept disambiguation and the hierarchy of the tree. The expansion process is therefore achieved by directly adding new weighted keywords to the initial query, which are extracted from this tree.

Experimental results using the precision metric at 5-points (5, 10, 15, 20 and 100) and Mean Average Precision (MAP) showed; a decrease in performance when queries are blindly extended by concepts, this is mainly due to the weak semantic coverage of the Arabic language by AWN, however, a positive enhancement in retrieving performance around 10% of MAP after application of the proposed approach.

Again, these results give interesting leads for future works on Arabic language using semantic trees to capture information needs of users.

REFERENCES

Abbache, A., Meziane, F., Belalem, G., & Belkredim, F. Z. (2018). Arabic query expansion using wordnet and association rules. In Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications (pp. 1239–1254). IGI Global. doi:10.4018/978-1-5225-5191-1.ch054

Abderrahim, M. E. A. (2014). *Concept based vs. pseudo relevance feedback performance evaluation for information retrieval system.* ArXiv Preprint ArXiv:1403.4362.

Aggarwal, N., & Buitelaar, P. (2012). *Query expansion using wikipedia and DBpedia. CLEF*. Online Working Notes/Labs/Workshop.

Aklouche, B., Bounhas, I., & Slimani, Y. (2019). Pseudo-relevance feedback based on locally-built co-occurrence graphs. *European Conference on Advances in Databases and Information Systems*, 105–119. doi:10.1007/978-3-030-28730-6_7

Arabic-Corpora. (2010). Retrieved April 7, 2019, from https://sourceforge.net/projects/ar-text-mining/files/Arabic-Corpora/

Aseervatham, S. (2009). A concept vector space model for semantic kernels. *International Journal of Artificial Intelligence Tools*, *18*(02), 239–272. doi:10.1142/S0218213009000123

Baziz, M., Boughanem, M., & Traboulsi, S. (2005). A concept-based approach for indexing documents in IR. *INFORSID*, *2005*, 489–504.

Bhogal, J., Macfarlane, A., & Smith, P. (2007). A review of ontology based query expansion. *Information Processing & Management*, *43*(4), 866–886. doi:10.1016/j.ipm.2006.09.003

Boubekeur, F., Boughanem, M., & Tamine-Lechani, L. (2007). Semantic information retrieval based on CP-nets. *2007 IEEE International Fuzzy Systems Conference*, 1–7.

Bounhas, I., Soudani, N., & Slimani, Y. (2020). Building a morpho-semantic knowledge graph for Arabic information retrieval. *Information Processing & Management*, *57*(6), 102124. doi:10.1016/j.ipm.2019.102124

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, *44*(1), 1–50. doi:10.1145/2071389.2071390

Colace, F., De Santo, M., Greco, L., & Napoletano, P. (2015). Weighted Word Pairs for query expansion. *Information Processing & Management*, *51*(1), 179–193. doi:10.1016/j.ipm.2014.07.004

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2004). A framework for selective query expansion. *International Conference on Information and Knowledge Management, Proceedings*, (3), 236–237. doi:10.1145/1031171.1031220

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems*, *29*(2), 1–34. doi:10.1145/1961209.1961211

El-khair, I. A. (2006). Effects of Stop Words Elimination for Arabic Information Retrieval : A Comparative Study. *International Journal of Computing & Information Sciences*, *4*(3), 119–133.

El Mahdaouy, A., El Alaoui, S. O., & Gaussier, E. (2018). Improving Arabic information retrieval using word embedding similarities. *International Journal of Speech Technology*, *21*(1), 121–136. doi:10.1007/s10772-018-9492-y

Elkateb, S., Black, W., Vossen, P., Rodríguez, H., Pease, A., Alkhalifa, M., & Fellbaum, C. (2006). Building a WordNet for Arabic. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, 29–34.

Habash, N. Y. (2010). Introduction to Arabic natural language processing. In Synthesis Lectures on Human Language Technologies (Vol. 3). doi:10.2200/S00277ED1V01Y201008HLT010

Hu, J., Deng, W., & Guo, J. (2006). Improving retrieval performance by global analysis. *18th International Conference on Pattern Recognition (ICPR'06)*, *2*, 703–706.

Huang, G., Wang, S., & Zhang, X. (2011). Query expansion based on associated semantic space. *Journal of Computers*, *6*(2), 172–177. doi:10.4304/jcp.6.2.172-177

Jones, K. S. (1971). *Automatic Keyword Classification for Information Retrieval*. Archon Books.

Karisani, P., Rahgozar, M., & Oroumchian, F. (2016). A query term re-weighting approach using document similarity. *Information Processing & Management*, *52*(3), 478–489. doi:10.1016/j.ipm.2015.09.002

Khatib, A. S. (1997). Terminological specifications and applications in the Arabic language. Cultural Fifteenth Season of the Arabic Language Academy of Jordan, 177–213.

Khoja, S., & Garside, R. (1999). *Stemming arabic text*. Computing Department, Lancaster University.

Khoury, R. (2011). Query classification using Wikipedia. *International Journal of Intelligent Information and Database Systems*, *5*(2), 143–163. doi:10.1504/IJIIDS.2011.038969

Li, Y., Luk, W. P. R., Ho, K. S. E., & Chung, F. L. K. (2007). Improving weak ad-hoc queries using wikipedia asexternal corpus. *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*, 797–798. doi:10.1145/1277741.1277914

list-of-queries. (2019). Retrieved April 7, 2019, from https://sourceforge.net/projects/queries-for-osac-arabic-corpus/files/

Liu, L., Cao, C., Zhang, C., & Tian, G. (2009). Sense recognition research of hyponymy based on concept space. *Chinese Journal of Computers*, *32*(8), 1651–1659.

Lucene. (2019). Retrieved April 7, 2019, from http://lucene.apache.org/

Mahgoub, A., Rashwan, M., Raafat, H., Zahran, M., & Fayek, M. (2014). Semantic query expansion for Arabic information retrieval. *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, 87–92.

Mallat, S., Zouaghi, A., Hkiri, E., & Zrigui, M. (2013). Method of lexical enrichment in information retrieval system in Arabic. *International Journal of Information Retrieval Research*, *3*(4), 35–51.

Mazari, A. C., Aliane, H., & Alimazighi, Z. (2013). A conceptual indexing approach for Arabic texts. *2013 ACS International Conference on Computer Systems and Applications (AICCSA)*, 1.

Pal, D., Mitra, M., & Datta, K. (2014). Improving query expansion using WordNet. *Journal of the Association for Information Science and Technology*, *65*(12), 2469–2478. https://doi.org/10.1002/asi.23143

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. In *Foundations and Trends in Information Retrieval* (Vol. 3). 10.1561/1500000019

Saad, M., & Ashour, W. (2010). OSAC: Open Source Arabic Corpora. *6th International Conference on Electrical and Computer Systems (EECS'10)*, 118–123.

Shaalan, K., Al-Sheikh, S., & Oroumchian, F. (2012). Query expansion based-on similarity of terms for improving Arabic information retrieval. *International Conference on Intelligent Information Processing*, 167–176.

Stairmand, M. A. (1997). Textual context analysis for information retrieval. *SIGIR Forum, 31*(1), 140–147. 10.1145/278459.258552

Xu, J., & Croft, W. B. (1996). Query expansion using local and global document analysis. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Xu, J., & Croft, W. B. (2000). Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, *18*(1), 79–112.

Ahmed Cherif Mazari was born in Berrouaghia City in Algeria, Magister from Algiers University in 2009, Engineer degree in computer science from National Institute for Computer Science at Algiers in 1996, Assistant Professor in the Department of Computer science at the University of Medea, researcher at the Laboratory of Advanced Electronic Systems (LSEA), Medea, Algeria. Currently a PhD student at the university of Biskra. His research interests are in Natural Language Processing, Information Retrieval, Sentiment Analysis and Deep Learning.

Djeffal Abdelhamid was born in Biskra City at Algeria, PhD from Biskra University in 2012, Master degree in image processing and AI from Biskra University in 2004, engineer degree in computer science from National Institute for Computer Science at Algiers in 1997, Assistant Professor in the Department of Computer science in the University of Biskra since December 2004, Member of LESIA laboratory and research team in image processing and satellite images since January 2005.