Arabic Query Expansion Using WordNet and Association Rules

Abbache, A, Meziane, F, Belalem, G and Belkredim, F Z

http://dx.doi.org/10.4018/IJIIT.2016070104

Table of Contents

International Journal of Intelligent Information Technologies

Volume 12 • Issue 3 • July-September-2016 • ISSN: 1548-3657 • eISSN: 1548-3665

An official publication of the Information Resources Management Association

Research Articles

1 **Using Belief Functions in Software Agents to Test the Strength of Application Controls: A Conceptual Framework**

Robert A. Nehmer, Oakland University, Rochester, MI, USA
Rajendra P. Srivastava, niversity of Kansas, Lawrence, KS, USA

20 **Customer Choice of Super Markets using Fuzzy Rough Set on Two Universal Sets and Radial Basis Function Neural Network**

A. Anitha, VIT University, Vellore, India
Debi Prasanna Acharjya, VIT University, Vellore, India

38 **A Neuro-Fuzzy Rule-Based Classifier Using Important Features and Top Linguistic Features**

Saroj Kr. Biswas, National Institute of Technology, Silchar, India
Monali Bordoloi, National Institute of Technology, Silchar, India
Heisnam Rohen Singh, National Institute of Technology, Silchar, India
Biswajit Purkayastha, National Institute of Technology, Silchar, India

51 **Arabic Query Expansion Using WordNet and Association Rules**

Ahmed Abbache, Department of Computer Science, University of Oran 1, Ahmed Ben Bella, Oran, Algeria
Farid Meziane, Informatics Research Centre, University of Salford, Salford, UK
Ghalem Belalem, Department of Computer Science, University of Oran 1, Ahmed Ben Bella, Oran, Algeria
Fatma Zohra Belkredim, Department of Computer Science, University Hassiba Ben Bouali of Chlef, Ouled Fares, Algeria

The *International Journal of Intelligent Information Technologies* is indexed or listed in the following: ACM Digital Library; Australian Business Deans Council (ABDC); Bacon's Media Directory; Burrelle's Media Directory; Cabell's Directories; Compendex (Elsevier Engineering Index); CSA Illumina; DBLP; DEST Register of Refereed Journals; Gale Directory of Publications & Broadcast Media; GetCited; Google Scholar; INSPEC; JournalTOCs; Library & Information Science Abstracts (LISA); MediaFinder; Norwegian Social Science Data Services (NSD); SCOPUS; The Index of Information Systems Journals; The Standard Periodical Directory; Thomson Reuters; Ulrich's Periodicals Directory; Web of Science

International Journal of Intelligent Information Technologies
Volume 12 • Issue 3 • July-September 2016

Arabic Query Expansion Using WordNet and Association Rules

Ahmed Abbache, Department of Computer Science, University of Oran 1, Ahmed Ben Bella, Oran, Algeria
Farid Meziane, Informatics Research Centre, University of Salford, Salford, UK
Ghalem Belalem, Department of Computer Science, University of Oran 1, Ahmed Ben Bella, Oran, Algeria
Fatma Zohra Belkredim, Department of Computer Science, University Hassiba Ben Bouali of Chlef, Ouled Fares, Algeria

ABSTRACT

Query expansion is the process of adding additional relevant terms to the original queries to improve the performance of information retrieval systems. However, previous studies showed that automatic query expansion using WordNet do not lead to an improvement in the performance. One of the main challenges of query expansion is the selection of appropriate terms. In this paper, the authors review this problem using Arabic WordNet and Association Rules within the context of Arabic Language. The results obtained confirmed that with an appropriate selection method, the authors are able to exploit Arabic WordNet to improve the retrieval performance. Their empirical results on a sub-corpus from the Xinhua collection showed that their automatic selection method has achieved a significant performance improvement in terms of MAP and recall and a better precision with the first top retrieved documents.

KEYWORDS

Arabic WordNet, Association Rules, Automatic Query Expansion, Information Retrieval, MAP

INTRODUCTION AND MOTIVATION

Information Retrieval (IR) is concerned with organizing, storing, retrieving and displaying information. IR systems aim to provide the user with an easy access to the information he/she is interested in. Usually the user is required to formulate his information need through a query; the IR system then provides the user with the relevant information in return (Baeza-Yates & Ribeiro-Neto, 1999). While interacting with the users, IR systems face many challenges, one of these being the vocabulary problem also referred to as vocabulary mismatch (Carpineto & Romano, 2012). To address this issue, researcher in the IR field proposed many solutions and the latest being the Automatic Query Expansion (AQE). This technique aims at reformulating the original query by adding new terms into it to achieve a better accuracy for the IR system. Various AQE techniques have been proposed and Cui et al. (2002) split them into two major classes: global analysis and local analysis.

The global analysis approaches are independent from the initial query or its result. Generally, they use external knowledge sources to select terms for expansion such as thesaurus or WordNet. Local analysis approaches formulate a new query on the basis of some retrieved documents of a previous search, for example relevance feedback (Bilel et al, 2011).

Adding new terms to the initial Query can take place prior to either the initial search or the relevance-feedback search (Cuna et al., 1992). The selection of these terms is a key phase in the IR process. There are several sources for terms selection, WordNet has been recognized as an important

DOI: 10.4018/IJIIT.2016070104

source of selection for query expansion. It is one of the largest and most widely used in the tasks of natural language processing (NLP), counting Word Sense Disambiguation (WSD) and Question Answering Systems (QAS) (Tingting et al., 1992).

Arabic is a vocalized language. It requires the adding of signs to the consonants to precisely define the pronunciation of a word. Hence, the non-vocalized Arabic word may have several possible meanings. Unfortunately, texts in Arabic languages, mainly Modern Standard Arabic (MSA) are not vocalized. For example, the non-vocalized word: (عالم) may mean by way of its vocalization: scientist (عَالِم) or world (عَالَم). This phenomenon makes the selection of appropriate synonyms for expansion more difficult in Arabic, a problem that is not faced by other languages.

Association rules have been used in several areas, including clustering and IR (Picariello & Rinaldi, 2007; Veeramalai & Kannan, 2011). In AQE, they have been used to provide semantic links between terms. In a previous study (Abbache et al., 2014), we have shown that AQE does not improve retrieval; but on the other hand, Interactive Query Expansion (IQE) improves retrieval. We concluded that if we can find a way to select appropriate terms from the Arabic WordNet instead of taking all the returned terms, we may achieve better results. In this study, we investigate the possibility of using association rules between terms based on the assumption that words in documents that associate with a word in the query are more likely to be related to that query word.

The remaining of the paper is organized as follows. Section two summarizes some related and similar work highlighting specifically the methodology used and the results obtained; Section three attempts to highlight the source of terms selection (Arabic WordNet). Section four presents the proposed technique for automatic query expansion. Section five describes the experiments and section six summarizes the conclusions.

RELATED WORK

Various methods and techniques have been proposed or used for AQE and these have been extensively studied in IR. Query expansion is the process of adding additional terms to the original query to improve the IR systems' performance. AQE is not new; in fact, it has been mentioned in earlier 1960 but has not reached maturity until very recently (Carpineto & Romano, 2012).

In an early work, Cuna et al. (1992) evaluated three different approaches in information retrieval systems to expand user queries. The basic unit used was a stem rather than word, based on the assumption that terms occurring together in the same document are more likely to be related. This was their first approach and they named it term co-occurrence. Their second approach was based on the concept of Soundex code, where they assigned the same code to terms that sound the same. String similarity was the idea of their last approach known nowadays as n-gram. Their experiment was performed on small scale collections: 26,280 records from the input of the Library and Information Science Abstracts database (1982-1985) and 114 queries were used. Their results indicated that the expansion methods used do not increase the search performance, and there was no significant difference between them.

As the volume of data increased, research on AQE has been substantially revised, and the topic has received more attention in recent literature on IR (Carpineto & Romano, 2012). Believing in the idea that since words which are located in proximity are semantically related, so the distance between them can be used to indicate their association. Vechtomova & Wang (2006) evaluated different distance functions to select query expansion terms. The experiment was performed using a large corpus of the TREC collection (Financial Times and Los Angeles Times) and Okapi as an IR system. Their method shows significant performance improvement compared to the original study that did not include expansion. They mentioned that the use of the number of relevant documents to extract the association between pair of word may improve the selection of query terms.

Han & Chen (2009) proposed a hybrid method for query expansion which combines two methods: ontology-based and neural networks. The first one has been applied to find similar users using a

collaborative filtering to explore semantic relationships. The second method aims to obtain terms from the most relevant web documents for user's queries. To test this new method, they collected ontology information from 251 Web pages (personnel information). They could found 18 similar users, and then a set of term was extracted from some relevant web documents from these similar users' research interests using Google search engine. They confirmed that this method could improve precision, and it required only few items of query information provided by the users.

In the last few years, the number of AQE techniques has increased significantly. Researchers have used a variety of approaches, applying several data sources to expand the user's query linked to the appropriate terms. In 2010, Xiangming & Kun (2010) performed a series of experiment using the TREC 2006 track Genomic data set. The aim of this experiment was to see how the effectiveness of an IR can be affected by using different query expansion strategies with variation in the average number of terms in a query, which they refer to as query complexity. The Unified Medical Language System (UMLS) was used as a source of term selection, with different indices being provided, word and string indexing. The results showed that expansion using string index perform better than expansion using word index. In terms of query complexity, they concluded that the expansion with short queries outperformed long queries. Finally, they proposed the use of the string index with short queries.

Vitaly & Yannis (2011) compared two measures between terms, the EWC semantic relatedness measure and Wikipedia-based Explicit Semantic Analysis (ESA) measure. Using Terrier[1] as an open source search engine, these two measures were used as a source of selection in a query expansion technique applied to the NTCIR as a test collection (187,000 articles presented at scientific conferences hosted by the Japanese academic societies). The authors indicated that the first measure outperformed ESA.

Based on the characteristic of the semantic web and ontology technology, Wang & Qiang (2012) proposed a combination of a global analysis and an ontology which first annotates the web documents, and then the terms in these documents are associated with the ontology concepts. Then the expansion is based on this relatedness; in their experimentation they use two datasets, the Reuter's Corpus Volume 1 (804414 English language news stories) and the annotated metadata ACM digital library (29030). Using two ontologies: WordNet and ACMCCS98, they confirmed that their method could improve the precision effectively. They noticed that the dataset scale and the ontology characteristic are two factors may affect the precision at the first 20 returned results.

In a new query expansion method, Francesco et al. (2013) proposed the use of a structured representation with named, weighted word pairs to expand the initial query. This method adopts the explicit relevance feedback to perform a new query. These authors extracted automatically the Weighted Word Pairs representation from documents using term extraction. TREC-8 dataset was used to test the model, allied to about 520,000 documents on 50 topics. The result confirmed that their approach retrieved more relevant documents than an approach based on a list of words only.

A real evaluation of query reformulation was performed by Abderrahim (2013) using external resources in an Arabic Information Retrieval System (IRS). Two different resources were used, the first was Arabic WordNet, and the second was the Arabic Dictionary of Meanings (ADM). They used a corpus of 22.000 Arabic documents and 50 queries. Lucene's APIs were used for indexing and searching. They developed four strategies: the simple search, blind, controlled, and weighted search. Every strategy is tested two times, with AWN and ADM. Experiments show that their approach performs an Arabic IRS with six per cent.

Pragati & Narendra (2014) showed the limits of pseudo relevance feedback (PRF) technique in query expansion, and proposed a new method which combines the standard use of PRF with a genetic fuzzy approach, and a semantic similarity notion. To retrieve the initial set of documents they used the Okapi similarity measure, and to select a candidate terms list they used the Jaccard similarity measure. Then they selected an optimal subset from this list using a genetic fuzzy approach. Furthermore, they observed that some terms from this subset are general and not related to the query; they decided to use a semantic similarity between this subset of terms and the original query terms in order to get the

appropriate set of the terms. The experiments were performed using the CISI dataset (1460 abstracts and 112 queries) and they conclude that their method improve the performance of query expansion in term of recall and precision.

Our own work differs from the related work in the following aspects. First, we focus on a particular type of queries, namely short queries. Second, our approach focuses on the use of the association rules between terms to select the appropriate query term's synonyms from Arabic WordNet only. Third, we show how an automatic query expansion using Arabic WordNet and association rules can improve the effectiveness of an IRS for the Arabic language.

ARABIC WORDNET

In this section, we attempt to explore and understand the structure of WordNet, how it is used and for what applications it is used. We will conclude by identifying its strengths and weaknesses. WordNet (first created for the English language) was based on the idea of using synonym sets (Picariello & Rinaldi, 2007) to represent lexical concepts described by the lexical matrix, mapping between word forms and word meanings. It was considered as a program ancillary to the machine-readable dictionary, allowing users to explore an online dictionary on the basis of semantic similarities rather than the alphabetical ones. Then it evolved from a dictionary browser into a self-contained lexical database (Fellbaum, 1998; Miller et al., 1990).

Arabic WordNet (AWN) (started in 2006) is a free and open source lexical database for Modern Standard Arabic (MSA). It follows the conception and the methodology of the Princeton WordNet (Elkateb et al., 2006) and EuroWordNet (Black et al., 2006) (Created for European languages). It groups Arabic words (Nouns, Verbs, Adjectives…) into sets of synonyms called synsets. Every synset can be thought of as representing a unique word sense (meaning or concept) and records the various semantic relations between them (Habash, 2010). It currently counts 11,269 synsets (7,960 names, 2,538 verbs, 110 adjectives, 661 adverbs), and 23,481 words (Abderrahim et al., 2013). In addition to synonym relation, WordNet includes other semantic relations such as hyponymy (a word whose semantic range is included in another word) and hypernymy (a word whose semantic range includes another word) (Habash, 2010). For example, a set like {(كتاب, book), (رسالة خطية, written message), ((رسالة, message), (خطاب, speech)} form an Arabic synset, because the elements of this set can be used to share the same approximate meaning. Consequently, synset scan be related and linked to each other by semantic relations, such as synonymy, antonymy, hyponymy, etc.

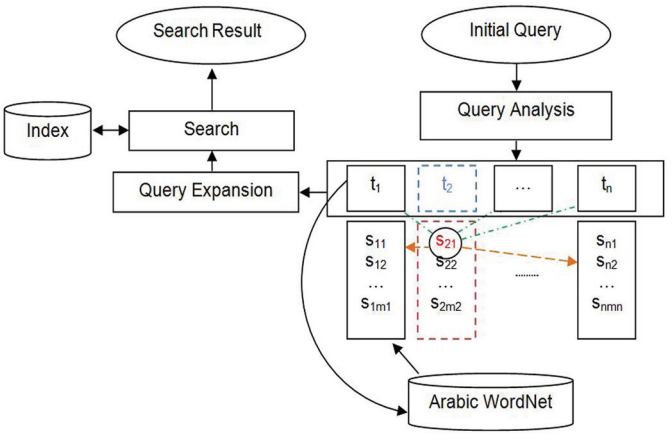
THE PROPOSED METHODOLOGY

In this section, we present our query expansion method based on Arabic WordNet and Association Rules for mining additional query terms. Figure 1 summarises the proposed query expansion method. The user represents usually his needs with a set of words which forms his initial query. To expand this query with appropriate words; we propose a methodology with several essential steps including a pre-processing phase. In the pre-processing phase, stop words (Frequently used words that do not help distinguish one document from the other as they do not add any semantic meaning to the sentence such as "the", "of" and "that") are removed to prepare the query for the next step which is the extraction and selection of synonyms using Arabic WordNet as the source for term selection.

In the literature, there are several sources for term selection designed to capture the vocabulary of a language such as TemaTres[2], Moby Thesaurus[3] and others. WordNet is also designed for the same purpose but it is the largest and most comprehensive resource that is developed and maintained manually by a global community.

The extraction of the useful tokens is performed by the analyser (Query Analysis). In our system, the analyser uses the light-stemming algorithm as specified in (Larkey et al., 2005). This algorithm was

Figure 1. The proposed query expansion method



compared with several stemmers based on morphological analysis and showed a better performance (Larkey et al., 2005).

In our approach we apply two query expansion methods in sequence to reformulate the query. We first apply the Interactive Query Expansion (IQE) and then the Automatic Query Expansion (AQE).

Interactive Query Expansion

In this method, the system provides several suggestions to the user, and the query reformulation is based on his selection. The user enters a set of keywords to the system which are analysed by the "Analyser" and the terms are then passed to the "Index Searcher" to return the search results. The "Synonyms Extractor" provides the user with different parts of speech (PoS) for each term in the original query. Based on the user selection, a list of suggestions is proposed. The user selects the appropriate synonyms, and starts the search again. For more details about this method we suggest to read our previous work (Abbache et al., 2014).

Automatic Query Expansion

AQE has the same steps as IQE; the main difference is that in AQE, the system rather than the user selects the appropriate synonyms. In this method, to select the appropriate synonyms we attempt to find if there is an association between the other terms of the query and their synonyms. Our approach to the selection of the query expansion terms is based on the assumption that words that tend to occur together in documents are likely to have similar, or related, meanings and hence can be associated. This association can occur between the synonym and one of the terms of the query or their synonyms.

To illustrate the problem better, let us make the following assumptions:

- Let $Q = \{t_1, t_2 \ldots t_n\}$ Where Q is a query that is analysed and decomposed into a set of terms;
- t_i (i=1 … n) is the i^{th} term in the query Q and
- n is the number of query terms;
- $S_i = \{S_{i1}, S_{i2} \ldots S_{imi}\}$ is the list of the synonyms of the term t_i where S_{ij} (j=1 … mi) is the j^{th} synonym in S_i and mi is the number of synonyms in S_i. We note here that the set S_i could be empty.
- We define a function *f* that calculates the association between terms where $f(t_i, t_j) = w$ is the weight or the association between the terms t_i and t_j. If $w = 0$ this mean that there is no association between the terms t_i and t_j.

Let us consider the synonym S_{11}. To select S_{11} as an appropriate term for expansion, it must associate with at least one term from the set $\{t_2 \ldots t_n\}$ or a term from the set of sets $\{S_2 \ldots S_n\}$. In addition to the selection of terms, it is common to also associate weights that reflect the importance of each term. In our method, we associate with the original query's terms the weight 1 and we associate the max weight returned by the function $f(t_i, t_j)$ to the selected synonyms.

The function f was developed using R which is a free software environment for statistical computing and graphics. Term-document matrix that describes the frequency of terms that occur in our collection of documents was created first, and from the association matrix extracted from this matrix we found which terms are highly associated with a query term. After a set of experiments, we conclude that the correlation between words in our collection of documents should be greater or equal to 0.3.

The following algorithm illustrates the detailed mechanism of our query expansion method:

EXPERIMENTS

Experiments were performed using Lucene APIs; an extremely rich and powerful full-text search library written in Java, supported by the Apache Software Foundation and released under the Apache Software License (Otis, et al., 2010). Retrieval systems contain two main phases; the indexing phase, which is responsible for indexing the collection to build an index, and the searching phase, which is responsible for searching this index. The first step in implementing our full-text searching application with Lucene was to build the index; its creation is the central component of the indexing process. To evaluate the expansion methods, we have segmented our experimentation into four search types that will be studied individually to estimate the augmentation of each type to improve the search performance. These search types are:

Figure 2. Algorithm of synonyms selection

```
Algorithm Synonyms Selection
Begin
    For all query terms do
        Get synonyms of current term
        For all synonyms of current term do
            Compare current synonym with other query terms
            If current synonym associate with other query terms then
                    Select synonym;
                    Weight = 1;
            Else
                    Compare current synonym with the synonyms of other query terms
                    If current synonym associate with synonym of other query terms then
                            Select synonym;
                            Weight = maximum association score;
                    EndIf;
            EndIf;
        EndFor;
    EndFor;
End.
```

- Simple Search (R1): Searching Lucene Index with no expansion.
- Automatic Query Expansion with Synonyms (R2): Searching Lucene Index by expanding user's queries using synonym relation.
- Automatic Query Expansion (RS): using WordNet and Association Rules.
- Interactive Query Expansion (RI): After several suggestions provided to the user, searching Lucene Index by expanding the query with his decision.

In the rest of this section we will introduce the dataset used and the performance measures that are used in this experimental study. The findings and the analysis of results are then discussed.

Test Collection

Lucene was used to index the Xnh-4500[4] Arabic corpus, a sub-corpora from the Xinhua[5] collection relating to 2008-2009 period (Brahmi et al., 2012). This subset contains 4,500 Arabic documents, the average number of words per document is 205.3 words, and the sub-collection contains 956,705 words with 152,601 distinct words. Each text document belongs to one of eight categories as shown in Table 1. We also used a set of short keywords queries (max three words) in various fields for evaluation, for example: "امتداد المرض", "بطولة كأس العالم".

Performance Measures

To evaluate the effectiveness of each strategy, the following performance measures are used:

- Recall: is the fraction of the documents that is relevant to the query that is successfully retrieved.
- Precision: is the fraction of retrieved documents that is relevant to the user's information need.
- F-measure: the weighted harmonic mean of precision and recall, the traditional F-measure is:

$$F = \frac{2.\text{precision}.\text{recall}}{\left(\text{precision} + \text{recall}\right)} \quad (1)$$

- Mean Average Precision: is the average precision across multiple queries/ranking.

Table 1. Documents categories

Domain	Doc. Number
China	563
Culture-Education	562
Economy	563
Middle-East	563
Science-Health	563
Sport	563
Tourism-Ecology	563
World	562
Total	4500

$$\mathrm{MAP} = \frac{\sum_{q=1}^{Q} \mathrm{AveP}(q)}{Q} \tag{2}$$

Results

Table 2 shows the different values for recall, precision; F-measure and MAP obtained by the system before and after the use of the different expansion strategies. A simple comparison of the results obtained, enables us to see that for any types of expansion there is an improvement in recall, and for RS and RI there is an improvement in MAP as well.

The recall in R2, RS and RI are respectively 27%, 10.5% and 18.2% which are better than the baseline. This means that there is an increase in the number of relevant documents retrieved by the system for each technique. In other words, QE using AWN with or without association rules improves the recall in Arabic Information Retrieval.

Unlike recall, there is a high decrease in precision using R2, and a smaller decrease using RS and RI. The precisions in R2, RS and RI are respectively 57%, 23.9% and 25.7% worse than the baseline. This means that there is a substantial increase in the number of irrelevant documents retrieved by the system. So the QE techniques used does not improve the precision in Arabic Information Retrieval. But our technique showed the lowest decrease at only 23.9%.

Taking into consideration both recall and precision, once again we see that our technique gives the lowest decrease. It can be observed clearly from Figure 3 that with regards to the MAP, our technique outperforms both the baseline and the other strategies.

Table 3 displays some results obtained by using the four strategies. Precision/recall are presented showing the interpolated average precision at eleven standard recall levels, using Lucene with and without query expansion.

The obtained results confirm that R2 is ineffective while RS and RI are. The benefits of our technique can be seen in Figure 4 which shows recall/precision graphs. We find that the performances of RS and RI are very close. The experimental results show that R2 does not bring improvement; RS improves retrieval in eight points of recall (0.0 to 0.7), and the performance degrades in the three points (0.8 to 1.0). RS improves retrieval in seven points of recall (0.0 to 0.6), and the performance degrades in four points (0.7 to 1.0). This means that the user gets relevant documents as well as irrelevant ones. The benefits of our method can be seen in Figure 4 which shows recall/precision curves.

However, it is stated by many researchers that what matters the most in an information retrieval system is the quality of the first results returned (Neumann, 2007). These are referred to as the top k retrieved documents (Christopher et al., 2008). Hence, we have recorded the results of the recall and precision when retrieving the first 20, 50, 100 and 150 documents using the four strategies. The results are summarised in Table 4 for recall and in Table 5 for the precision and graphically illustrated in Figure 5 and Figure 6 respectively.

With regards to recall, the global results have shown that our method improves on the baseline results but did not perform as well as R2 and RI. However, when considering only the top 20 results,

Table 2. The average recall, precision, F-measure and MAP obtained in different methods

Domain	R1	R2	RS	RI
Recall	0.70	0.88 (+27.0%)	0.77 (+10.5%)	0.82 (+18.2%)
Precision	0.21	0.09 (-57.0%)	0.16 (-23.9%)	0.15 (-25.7%)
FM	0.28	0.16 (-45.4%)	0.24 (-14.6%)	0.24 (-15.8%)
MAP	0.37	0.31 (-15.8%)	0.42(+15.8%)	0.41(+12.6%)

Figure 3. The average recall, precision, F-measure and MAP obtained in different methods

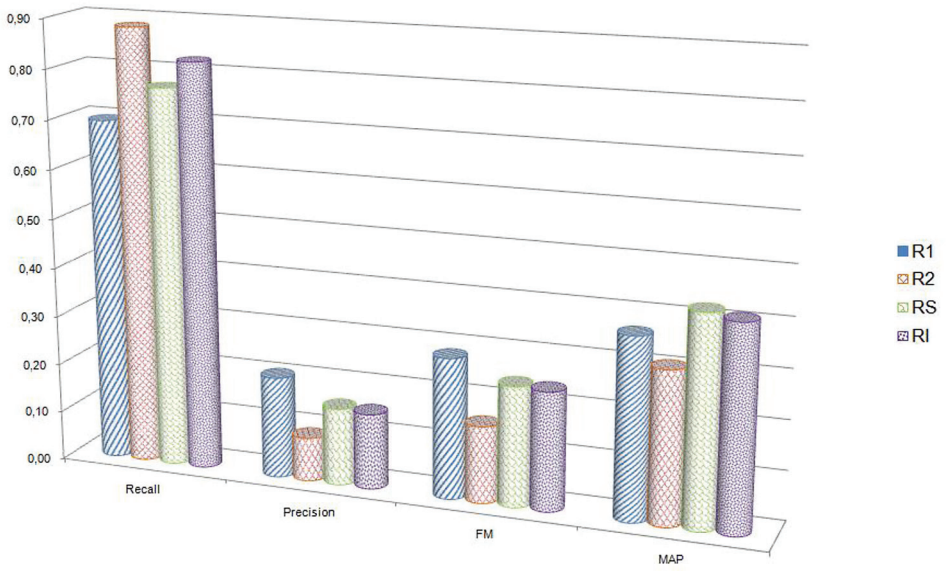


Table 3. The average precision obtained in different methods

Domain	R1	R2	RS	RI
0.0	0.84	0.82 (-02%)	0.88 (+05%)	0.86 (+02%)
0.1	0.55	0.56 (+00%)	0.74 (+34%)	0.68 (+23%)
0.2	0.45	0.43 (-03%)	0.58 (+29%)	0.54 (+21%)
0.3	0.42	0.39 (-08%)	0.48 (+13%)	0.49 (+16%)
0.4	0.38	0.34 (-10%)	0.42 (+11%)	0.44 (+17%)
0.5	0.36	0.33 (-08%)	0.37 (+04%)	0.40 (+13%)
0.6	0.30	0.26 (-13%)	0.34 (+13%)	0.35 (+16%)
0.7	0.30	0.23 (-22%)	0.28 (-05%)	0.31 (-03%)
0.8	0.29	0.19 (-33%)	0.24 (-17%)	0.28 (-04%)
0.9	0.27	0.13 (-50%)	0.22 (-18%)	0.22 (-16%)

our method improves the results for the baseline by 22.2% and outperforms the results of R2 (-22.2%) and RI (+11.1%). Our method also outperforms the other two methods when considering the first 100 documents. However, when considering the first 50 and 150 documents, our method outperforms the baseline and R2 but not RI.

As shown in Figure 5, R2 decrease recall in top 20, 50, 100 and 150 retrieved documents; while in the global results it showed the opposite. Meaning that the number of relevant documents retrieved by the system is small; In other words, R2 does not improves the recall in first top k retrieved document.

With regards to precision, globally the precision decrease for all the three methods used. However, when considering only the first 20 documents, our method improves the baseline by +38.6% and outperforms R2 (-11.4%) and RI (+18.2%). When considering the first 50, 100 and 150 documents, our method outperforms the baseline and R2 but not RI.

Figure 4. Recall/Precision Graph for the Xnh-4500 Test Collection

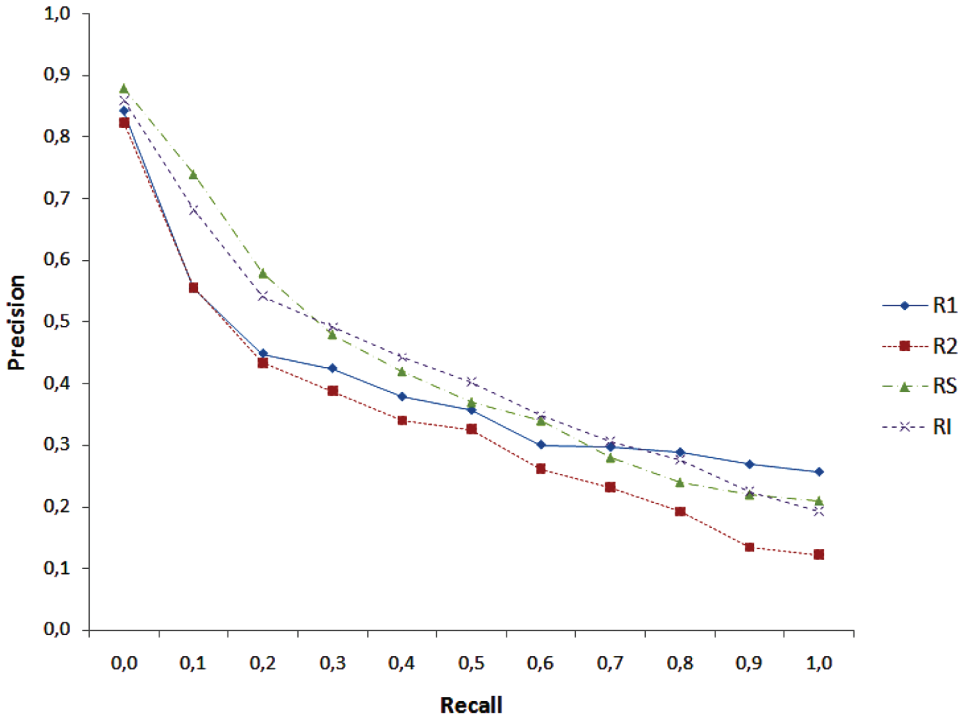


Table 4. The average Recall obtained in 20, 50, 100, 150 documents retrieved

		Number of Documents Retrieved			
		20	**50**	**100**	**150**
Recall	R1	0.09	0.18	0.29	0.41
	R2	0.07 (-22.2%)	0.16 (-11.1%)	0.24 (-17.2%)	0.34(-17.1%)
	RS	0.11 (+22.2%)	0.20 (+11.1%)	0.32 (+10.3%)	0.42 (+2.4%)
	RI	0.10 (+11.1%)	0.21 (+16.7%)	0.29 (0.0%)	0.44 (+7.3%)

Table 5. The average Precision obtained in 20, 50, 100, 150 top retrieved documents

		Number of Documents Retrieved			
		20	**50**	**100**	**150**
Precision	R1	0.44	0.38	0.33	0.29
	R2	0.39 (-11.4%)	0.38 (0.0%)	0.32 (-3.0%)	0.30 (+3.4%)
	RS	0.61 (+38.6%)	0.46 (+21.1%)	0.40 (+21.2%)	0.36 (+24.1%)
	RI	0.52 (+18.2%)	0.50 (+31.6%)	0.41 (+24.2%)	0.37 (+27.6%)

Figure 5. The average Recall obtained in 20, 50, 100, 150 documents retrieved

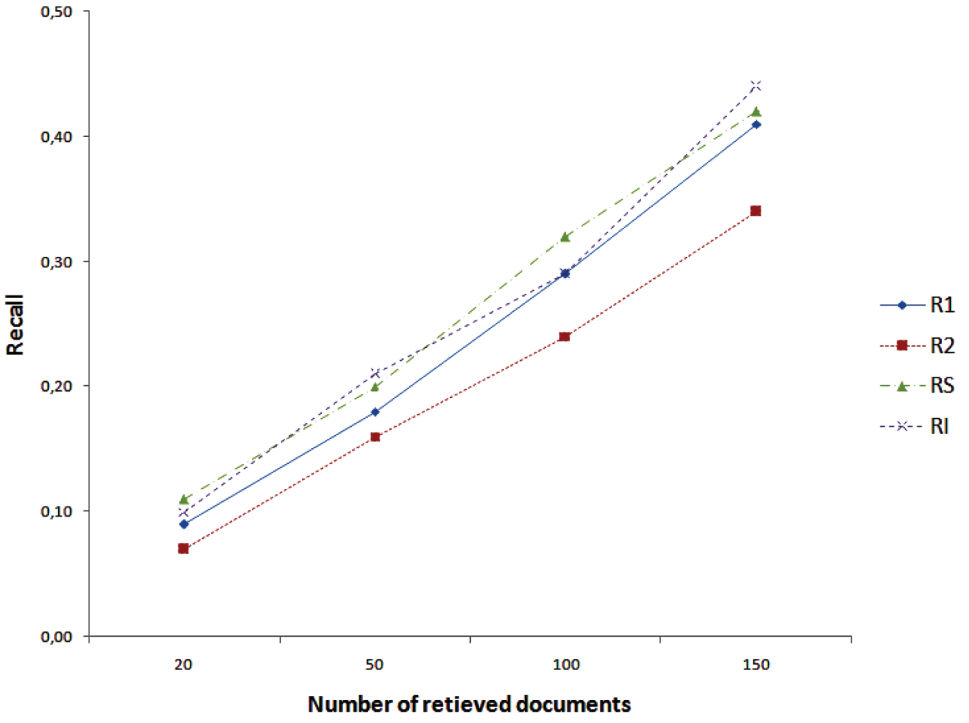
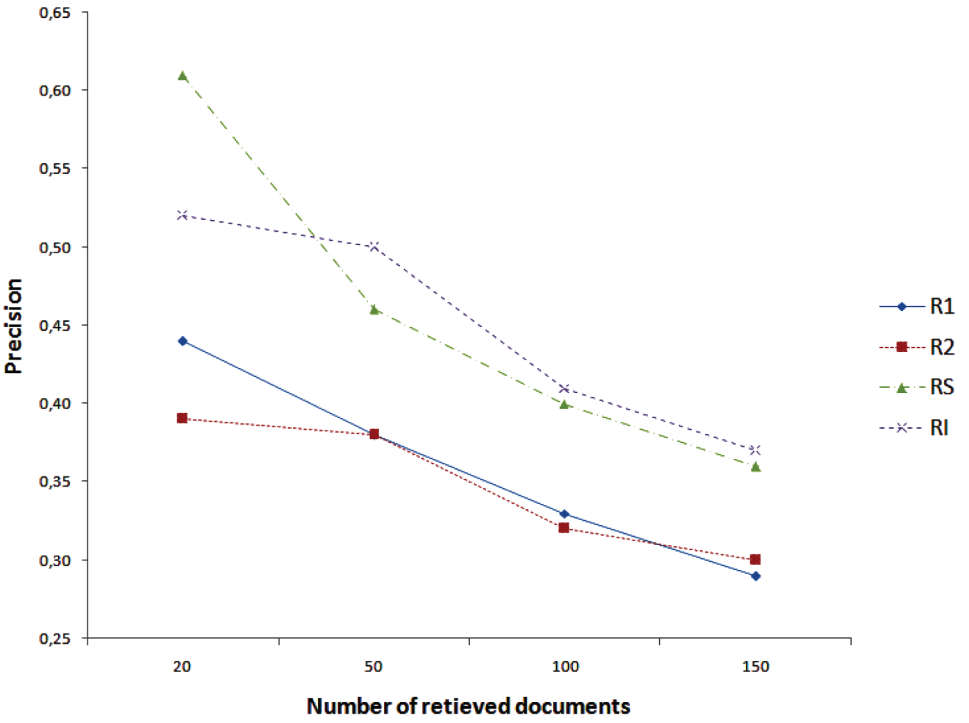


Figure 6. The average Precision graph obtained in 20, 50, 100, 150 top retrieved documents



The most important difference is the large decrease in the number of irrelevant documents in the first top 20, 50, 100 and 150 retrieved results. Figure 6 shows the substantial increase in precision for our strategy RS. Our selection method significantly improves precision in the first top k retrieved documents and prevents the user from receiving a large number of irrelevant information.

CONCLUSION

This work revealed the challenges and limitations when using AWN to expand user's queries when attempting to improve the effectiveness of Arabic information retrieval applications. To expand the query words, we used the synonym semantic relations included in AWN. We have performed a series of experiments to evaluate the effectiveness of three expansion strategies. One is the selection of all of a query term's synonyms; the second is to select synonyms using our method (automated selection based on AR) and the third one is the interactive approach where the terms are selected by the user. The experiments show that all strategies yield a significant increase in the recall measure, which means that the user gets more relevant documents compared to the basic search. On the other hand, there is a high decrease in precision using the first strategy and a little decrease using the second and the third. This means that the user gets more irrelevant documents, so there is a lot of 'noise' in the result returned to the user. Our proposed strategy and the interactive one improve the results of retrieval regarding the mean average precision, but the first strategy (R2) does not. This is due to the fact that the number of synonyms for each word is excessive. Adding all valid synonyms to a query introduces noise which degrades the performance. Also, AWN does not contain synonyms for all Arabic words, and some terms, such as proper names, are not included at all. Another problem is the polysemous nature of Arabic words. Our method also outperforms the baseline and the other strategies, and improves an IR system significantly with regard to the precision at the first top k retrieved documents. Nevertheless, the experiments allow us to conclude that with a good automatic selection of the right synonyms, the use of Arabic WordNet as a source of linguistic information for automatic query expansion improves the effectiveness of Arabic information retrieval.

NOTES

1. http://terrier.org/ Terrier is a flexible, efficient, and effective open source search engine.
2. http://www.vocabularyserver.com/
3. http://moby-thesaurus.org/
4. Link to this corpus: https://sites.google.com/site/abderrezakbrahmi/arabic-datasets
5. Chinese news agency with an edition online: http://arabic.news.cn/

ACKNOWLEDGMENTS

The authors warmly thank Pr. Christopher Bagley for valuable suggestions and proofreading, although the authors are solely responsible for omissions or errors in the final text.

REFERENCES

Abbache, A., Barigou, F., Belkredim, F. Z., & Belalem, G. (2014). The Use of Arabic WordNet in Arabic Information Retrieval. *International Journal of Information Retrieval Research*, *4*(3), 54–65.

Abderrahim, M. A., Abderrahim, M. E. A., & Chikh, M. A. (2013). Using Arabic WordNet for semantic indexation in information retrieval system. *International Journal of Computer Science Issues*, *10*(1), 327–332.

Abderrahim, M. E. A. (2013). Utilisation Des Ressources Externes Pour la Reformulation des Requêtes Dans un Système de Recherche d'Information. *The Prague Bulletin of Mathematical Linguistics*, *99*(1), 85–97.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. USA: Addison Wesley Longman Publishing.

Bhatnagar, P., & Pareek, N. (2014). Improving pseudo relevance feedback based query expansion using genetic fuzzy approach and semantic similarity notion. *Journal of Information Science*, *40*(4), 523–537. doi:10.1177/0165551514533771

Bilel, E., Ibrahim, B., Oussama, B. K., Fabrice, E., & Narjès, B. B. S. (2011). Towards a Possibilistic Information Retrieval System Using Semantic Query Expansion. *International Journal of Intelligent Information Technologies*, *7*(4), 1–25. doi:10.4018/jiit.2011100101

Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., & Fellbaum, C. (2006, January 22-26). Introducing the Arabic WordNet project.*Proceedings of the Third International WordNet Conference (GWC)*, Korea (pp. 255-299).

Brahmi, A., Ech-Cherif, A., & Benyettou, A. (2012). Arabic texts analysis for topic modeling evaluation. *Information Retrieval Journal*, *15*(1), 33–53. doi:10.1007/s10791-011-9171-y

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. *Journal ACM Computing Surveys*, 44(1).

Christopher, D. M., Prabhakar, R., & Hinrich, S. (2008). *Introduction to Information Retrieval*. England: Cambridge University Press.

Cui, H., Wen, J. R., Nie, J. Y., & Ma, W. Y. (2002, May 7-11). Probabilistic query expansion using query logs, *Proceedings of the Eleventh International World Wide Web Conference*, Honolulu, Hawaii, USA (pp. 325-332). ACM. doi:10.1145/511446.511489

Cuna, E. F., Alexander, M. R., & Peter, W. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. *Journal of Information Science*, *18*(2), 139–147. doi:10.1177/016555159201800208

Elkateb, S., Black, W., Vossen, P., Farwell, D., Rodríguez, H., Pease, A., & Alkhalifa, M. (2006, October 23). Arabic WordNet and the Challenges of Arabic.*Proceedings of Arabic NLP/MT Conference*, London, UK.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge: MIT Press.

Francesco, C., Massimo, De, S., Luca, G., & Paolo, N. (2013, January 16-17). A Query Expansion Method based on a Weighted Word Pairs Approach. *Proceedings of Fourth Italian Information Retrieval Workshop PISA*, Pisa, Italy (pp. 17-28).

Habash, N. Y. (2010). Introduction to Arabic natural language processing. *Machine Translation*, *24*(3-4), 285–289.

Han, L., & Chen, G. (2009). HQE: A hybrid method for query expansion. *Expert Systems with Applications*, *36*(4), 7985–7991. doi:10.1016/j.eswa.2008.10.060

Larkey, L., Ballesteros, L., & Connell, M. (2005). Light Stemming for Arabic Information Retrieval, Arabic Computational Morphology: Knowledge based and Empirical Methods. In A. Soudi, A. Van den Bosch, & G. Neumann (Eds.), *Text, Speech, and Language Technology* (pp. 221–243). Netherlands: Springer.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to Wordnet: An online lexical database. *International Journal Lexicography*, *3*(4), 235–244. doi:10.1093/ijl/3.4.235

Neumann, T. (2007, September 3-7). Optimizing ranked retrieval.*Proceedings of the 18th international conference on Database and Expert Systems Applications*, Regensburg, Germany (pp. 329-338). doi:10.1007/978-3-540-74469-6_33

Otis, G., Hatcher, E., & McCandless, M. (2010). *Lucene in Action* (2nd ed.). Stamford, USA: Manning Publications.

Picariello, A., & Rinaldi, A. M. (2007). User relevance feedback in semantic information retrieval. *International Journal of Intelligent Information Technologies*, *3*(2), 36–50. doi:10.4018/jiit.2007040103

Tingting, W., Yonghe, L., Huiyou, C., Qiang, Z., & Xianyu, B. (2014). A Semantic Approach for Text Clustering Using Wordnet and Lexical Chains. *Expert Systems with Applications*, *42*(4), 2264–2275.

Vechtomova, O., & Wang, Y. (2006). A study of the effect of term proximity on query expansion. *Journal of Information Science*, *2*(4), 324–333. doi:10.1177/0165551506065787

Veeramalai, S., & Kannan, A. (2011). Intelligent Information Retrieval Using Fuzzy Association Rule Classifier. *International Journal of Intelligent Information Technologies*, *7*, 14–27.

Vitaly, K., & Yannis, H. (2011). A Query Expansion Technique Using the EWC Semantic Relatedness Measure. *Informatica International journal of computing and informatics*, 35(4), 401-406.

Wang, Z., & Qiang, N. (2012). Research on Hybrid Query Expansion Algorithm. *International Journal of Hybrid Information Technology*, *5*(2), 207–212.

Xiangming, M., & Kun, L. (2010). Towards effective genomic information retrieval: The impact of query complexity and expansion strategies. *Journal of Information Science*, *36*(2), 194–208. doi:10.1177/0165551509357856

Call for Articles

International Journal of Intelligent Information Technologies

Volume 12 • Issue 3 • July-September 2016 • ISSN: 1548-3657 • eISSN: 1548-3665
An official publication of the Information Resources Management Association

MISSION

The advent of the World Wide Web has sparked renewed interest in the area of intelligent information technologies. There is a growing interest in developing intelligent technologies that enable users to accomplish complex tasks in web-centric environments with relative ease, utilizing such technologies as intelligent agents, distributed computing in heterogeneous environments, and computer supported collaborative work. The mission of the **International Journal of Intelligent Information Technologies (IJIIT)** is to bring together researchers in related fields such as information systems, distributed AI, intelligent agents, and collaborative work, to explore and discuss various aspects of design and development of intelligent technologies. This journal provides a forum for academics and practitioners to explore research issues related to not only the design, implementation and deployment of intelligent systems and technologies, but also economic issues and organizational impact. Papers related to all aspects of intelligent systems including theoretical work on agent and multi-agent systems as well as case studies offering insights into agent-based problem solving with empirical or simulation based evidence are welcome.

COVERAGE AND MAJOR TOPICS

The topics of interest in this journal include, but are not limited to:

Agent-based auction, contracting, negotiation, and ecommerce • Agent-based auction, contracting, negotiation, and e-commerce • Agent-based control and supply chain • Agent-based simulation and application integration • Cooperative and collaborative systems • Distributed intelligent systems and technologies • Human-agent interaction and experimental evaluation • Implementation, deployment, diffusion, and organizational impact • Integrating business intelligence from internal and external sources • Intelligent agent and multi-agent systems in various domains • Intelligent decision support systems • Intelligent information retrieval and business intelligence • Intelligent information systems development using design science principles • Intelligent Web mining and knowledge discovery systems • Manufacturing information systems • Models, architectures and behavior models for agent-oriented information systems • Multimedia information processing • Privacy, security, and trust issues • Reasoning, learning, and adaptive systems • Semantic Web, Web services, and ontologies

ALL INQUIRIES REGARDING IJIIT SHOULD BE DIRECTED TO THE ATTENTION OF:

Vijayan Sugumaran, Editor-in-Chief • IJIIT@igi-global.com

ALL MANUSCRIPT SUBMISSIONS TO IJIIT SHOULD BE SENT THROUGH THE ONLINE SUBMISSION SYSTEM:

http://www.igi-global.com/authorseditors/titlesubmission/newproject.aspx

IDEAS FOR SPECIAL THEME ISSUES MAY BE SUBMITTED TO THE EDITOR(S)-IN-CHIEF

PLEASE RECOMMEND THIS PUBLICATION TO YOUR LIBRARIAN

For a convenient easy-to-use library recommendation form, please visit:
http://www.igi-global.com/IJIIT