Received: March 26, 2018

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Improving Retrieval Performance Based on Query Expansion with Wikipedia and Text Mining Technique

Siham Jabri[1]* **Azzeddine Dahbi[1]** **Taoufiq Gadi[1]** **Abdelhak Bassir[1]**

[1] *Faculty of Science and Technology, Hassan 1st University Settat,*
Laboratory Informatics,Imaging and Modelling of Complex Systems, Morocco
* Corresponding author's Email: si.jabri@uhp.ac.ma

Abstract: Textual query is the simple mean for communicating with a retrieval system. However, there is a risk of providing an incomplete query which hinders the system from satisfying the user information needs. By reformulating the queries, query expansion is solution for this problem, this mainly relies on an accurate choice of the added terms to an initial query. It can yield a large number of irrelevant terms, which in turn negatively influences quality of retained documents. In this paper, we propose Query Expansion approach. It consists of reformulating queries by semantically related terms extracted from a semantic graph called query graph derived from Wikipedia. Furthermore, we propose a similarity measure which computes the similarity between a candidate terms and initial query using the query graph, Explicit Semantic Analysis (ESA) measure, and text mining technique. The experiments on Text Retrieval Conference (TREC) collection show that the proposed approach performs significantly better than the baseline system and some existing techniques.

Keywords: Query Expansion, Wikipedia, ESA, Query graph, Association rules, TREC.

1. Introduction

Since the beginning of the 21th century, the domain of technologies and computer science has known an accelerating development. As a result, big data was and still is a big challenge for the Natural language processing and information retrieval systems. It prevents people from searching information needs. One of reasons is that user queries to information retrieval systems are short and incomplete for describing and characterizing the relevant documents. Therefore, it seems natural to try to expand or reformulate the query by adding related terms that have not been explicitly mentioned by the user. Query expansion is an effective solution to reduce the usual query/document mismatch and improve retrieval performance. It not only increases the precision by putting the relevant documents at the top of results, but also the recall by retrieving relevant documents that cannot be retrieved by original query.

Query expansion is a long-standing research topic in information retrieval that has preoccupied researchers and there is a lot of studies that have been focus on it. For example, in reference [1] authors propose to integrate a term classification process to predict the usefulness of expansion terms, in reference [2] the idea of authors is to integrate the original query with feedback documents in a single probabilistic mixture model and regularize the estimation of the language model parameters, in reference [3] authors classify TREC topics into three categories based on Wikipedia: 1) entity queries 2) ambiguous queries and 3) broader queries and study the effectiveness of three methods for expansion term selection, in reference[4] Relevance feedback is an automatic process designed to produce improved query formulations following an initial retrieval operation, and in reference [5] authors expand short queries for micoblog retrieval by semantically related terms extracted from Wikipedia, DBpedia and unstructured texts using textmining techniques. However, despite all the researches, it continues to present a practical difficulty, so there is a need of some automatic techniques that can semantically reformulate and describe the user query.

In this paper, we propose a query expansion approach based on an external structured knowledge resource namely Wikipedia, Explicit semantic analysis (ESA) and association rules technique. The contributions of this work are as follows: First, we use the semantic interpretation ESA [6] for detecting the related terms to the query and building the expansion graph. Second, for avoiding the inclusion of non-similar terms in the extended queries, we propose a new semantic relatedness measure that combines an association rules technique [7], semantic measure [6] and the expansion graph.

The remainder of the paper is organized as follows: Section 2 presents related works and discusses the necessary scientific background about vector space model, association rules mining and explicit semantic analysis. Then, a detailed description of our approach is presented in Section 3. Experimental results and discussion are reported in Section 4. In Section 5, we conclude this paper and we present a future works.

2. Related works and background

2.1 Related works

A great deal of work has been done on query expansion and several approaches deal with the difficulty of providing a precise query to the retrieval system. Moreover, it has considered as an effective technique to improve the retrieval performance. Nevertheless, query expansion still suffers from limitations due to the fact that most of these techniques were based on a classical modelling which was developed to find the related terms to the query but not really the appropriate.

Most of query expansion approaches used a semantic resource to extract the appropriate expansion terms. For instance, Lv et al. [8] and Li et al. Authors in [9] used knowledge terms derived from the semantic resource called Freebase to expand the initial query. The authors explored Freebase to extract the global and local expansion keywords that are related to the original query and the pseudo-relevant documents ranked at the top of results.

Brandao [10] proposed two query expansion approaches. The first one is an unsupervised entity-oriented query expansion, which selects expansion terms using taxonomic features devised by the semantic structure. The second one is the involvement of machine learning techniques in order to select and rank the entities oriented for query expansion. El Ghali et al. [11] proposed a query expansion method for Web short queries using the Latent Semantic Analyses (LSA) technique which is based on the context around the query. This context is extracted from the search engine query logs by a three-query suggestion method: The Cosine Similarity, the Language Models, and their fusion. A context-aware query expansion method was another work of the same authors [12], the approach is based on LSA method and the user query is enriched by additional context extracted from the past user's queries Log. Anand and Kotov [13] used term graphs constructed from document collections such as encyclopedias (DBpedia) and knowledge bases (ConceptNet), as sources of semantically related terms for query expansion. Jain et al. [14] proposed a method that investigate the role of graph structure for query expansion and determine the importance of each node in the graph using an external resource called WordNet. The most important nodes representing word senses were identified and added to original query. Recently, Wikipedia has become an important external resource for Information retrieval. A several studies suggested to expand the original query using this resource. According to the reference [15], Boston et al. proposed a tool titled Wikimantic for disambiguating terms in search queries and for augmenting queries with expansion terms. By exploiting Wikipedia articles and their reference relations, this method defines an Atomic-Concept as a simple form of a concept for generating a set of terms. Zhao et al. [16] described a method for named entity disambiguation, which includes a query expansion based on Wikipedia terms through co-occurrence mentions. Two main strategies for identifying candidates are: 1) queries that contain abbreviations, a match is made to terms which have similar capitalization; 2) queries that contain continuous strings where the first letter of the string is also a capital letter, a match can be made to a candidate. The Wikipedia data used by this approach includes article titles, article content and article redirections. An initial query is used to retrieve the top k documents. Any candidate terms that are identified in the article collection become part of the collection of enhancement terms and articles returned become part of the article collection. This approach is titled feedback-query-expansion method, because it incorporates a feedback loop to find candidates during retrieval. Bruce et al. [17] uses Wikipedia and its hyperlink structures to find related terms for reformulating a query using link probability weighting and link-based measure by counting the number of documents where the term is already a hyperlink divided by the number of documents where the term appeared. This approach contains six steps outlined as: 1) The user's initial query is received, 2) aspects of the query are identified, 3) Wikipedia articles are selected, 4) aspect vocabulary is constructed, 5) finding under represented aspects, and 6) query expansion.

Our work is different from the previous works in two aspects. First, we introduce an expansion graph derived from Wikipedia based on Explicit Semantic Analysis (ESA) referenced in [6] for generating expansion terms. Second, relatedness between expansion terms and original query is using a new measure that combines a score based on the expansion graph, measure of an association rules technique based on multi-criteria optimization [18] and ESA.

2.2 Background

2.2.1. Vector space model for information retrieval

The basis in the vector space model is that the items in the information retrieval are represented as vectors in a vector space. The vector space model can be divided in to three stages. The first stage is the indexing where terms are extracted from the text document. The second stage is the weighting of the indexed terms and the last stage ranks the document with respect to the query according to a similarity measure.

Document indexing: It is evident that a stop words don't describe the content. Those words are removed from the document vector by using automatic document indexing, so the document will only be represented by significant words. The indexing is based on term frequency, where terms that have both high and low frequency within a document are considered to be function words [19,20,21] which are removed. The item vector space is a n-dimensional space, where *n* is the number of different terms used to index a set of documents. Document *i* (d_i) represented by a vector. Its magnitude in dimension *j* is w_{ij} where:

$$\begin{cases} w_{ij} > 0 & \text{if item j occurs in document i} \\ w_{ij} = 0 & \text{otherwise} \end{cases} \quad (1)$$

w_{ij} is the weight of item *j* in document *i*.

Term Weighting TF-IDF: The term weighting for the vector space has entirely been based on single term statistics. There are three main factors: term frequency factor, collection frequency factor and length normalization factor. These three factors are multiplied together to make the resulting term weight. The term frequency TF describes the document content and is generally used as the basis of a weighted document vector [22]. The concept of TF is that a term appearing many times within a document is likely to be more important than a term that appears only once.

$$TF_{ij} = \frac{f_{ij}}{l_i} \quad (2)$$

Where f_{ij} is the frequency of term *j* in document *i*, and l_i is the length of document *i*.

There are various weighting schemes to discriminate one document from others. This factor is called inverse document frequency *IDF*. It assumes that a term occurring in a few documents is likely to be a better discriminator than a term that appears in most or all documents [20].

$$IDF_i = \log(\frac{n}{n_j}) + 1 \qquad n_j > 0 \quad (3)$$

Where *n* is the number of documents and n_j is the number of documents in which term *j* occurs.

2.2.2. Explicit semantic analysis

Explicit Semantic Analysis, or ESA is a method for semantic representation of natural language texts. ESA is a process based on knowledge concepts explicitly defined by humans through Wikipedia.

The semantic of a given word is described by a vector presenting the word's association strengths to Wikipedia-derived concepts. A single Wikipedia article presents one concept and is defined as a vector of words that occur in this article weighted by the score *TF.IDF*. Once these concept vectors are generated, an inverted index is created to map back from each word to the concepts it is associated with. [6]. The ESA measure between a document vector *d* and query *q* is defined as follows:

$$ESA(dt, qt) = \frac{\overrightarrow{dc} \times \overrightarrow{qc}}{\left\|\overrightarrow{dc}\right\| \times \left\|\overrightarrow{qc}\right\|} \quad (4)$$

Where the numerator represents the dot product of the two concept vectors generated by ESA that represent document *dt* and the query *qt*. While the denominator represents the multiplication of the two concept vectors length.

2.2.3. Association rules mining

An association rule is a relation $T_1 \Rightarrow T_2$, where the itemsets T_1 and T_2 constitute respectively its premise and conclusion parts [7]. Thus, a rule presents a probability to have the terms of the conclusion in the transaction, given that those of the premise are already there, knowing that a transaction can be a document or a part of document.

Given an itemset *T*, the support of *T* is equal to the number of transactions in the documents collection containing all the terms of *T*. The support is formally defined as follows:

$$Supp(T) = \frac{\{d \in C; T \subseteq d\}}{|T|} \qquad (5)$$

Where *C* is the documents collection, *d* is a transaction ($d \in C$) and *T* is an itemset of the collection. Given a rule R: $T_1 \Rightarrow T_2$, the support of *R* is computed as follows:

$$Supp(R) = Supp(T_1 \, Y \, T_2) \qquad (6)$$

The confidence of *R* is computed as follows:

$$Conf(R) = \frac{Supp(T_1)}{Supp(T_1 \, Y \, T_2)} \qquad (7)$$

An association rule *R* is frequent if its support value, is greater than or equal to a user-defined threshold denoted minsupp, and the rule is said to be valid if the confidence value is greater than or equal to a user-defined threshold denoted *minconf*. There are other measures that have been proposed to select the interesting rules [18]:

3. The proposed approach

After presenting the related works and background. Thus, we will firstly describe our automated method of query expansion illustrated in Fig.1 and then discuss it in three parts which are terms extraction, query graph building and filtering.

Table 1. Some interesting measures

Measure	Formula
Lift	$\text{Lift}(R) = \frac{Supp(T_1 \cup T_2)}{Supp(T_1) \times Supp(T_2)}$
Information Gain	$GI(R) = \log_2(\frac{Supp(T_1 \cup T_2)}{Supp(T_1) \times Supp(T_2)})$
Jaccard	$JRD(R) = \frac{Supp(T_1 \cup T_2)}{Supp(T_1 \cup \overline{T_2}) \times Supp(T_2)}$
Cosinus	$Cos(R) = \frac{Supp(T_1 \cup T_2)}{\sqrt{Supp(T_1) \times Supp(T_2)}}$

The first part presented in Fig.1(a) concerns terms extraction from Wikipedia, a frequency according to TF.IDF is then computed for each term and only those having a frequency greater than a certain threshold are retained. At this stage of the process, each retained term could have several related terms and so on. The second one in Fig.1(b) concerns graph building. Given all the candidate terms to be ranked, we construct a graph in which each node represents a term, and each edge measures the relatedness between the two corresponding terms. Therefore, when deciding whether to select a term for query expansion in the third phase (Fig.1(c)), we consider two factors: first, this term is quite close to other terms that have high query-term relevance; and second, this term is strongly correlated with the query terms in an unstructured texts content or semantically related to the query terms using Wikipedia concepts.

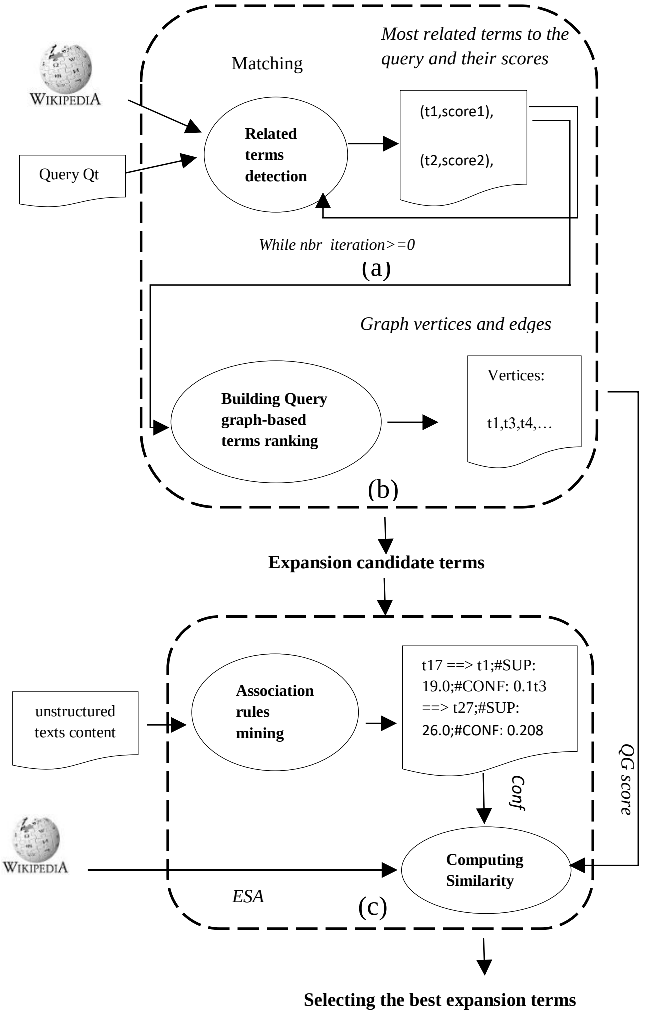
WIKIPEDIA
Matching
Most related terms to the query and their scores
Query Qt
Related terms detection
(t1,score1),
(t2,score2),
While nbr_iteration>=0
(a)
Graph vertices and edges
Building Query graph-based terms ranking
Vertices:
t1,t3,t4,…
(b)
Expansion candidate terms
unstructured texts content
Association rules mining
t17 ==> t1;#SUP: 19.0;#CONF: 0.1t3 ==> t27;#SUP: 26.0;#CONF: 0.208
Conf
QG score
WIKIPEDIA
ESA
Computing Similarity
(c)
Selecting the best expansion terms

Figure. 1 Description of the automated of query expansion

3.1 Generating candidate terms from Wikipedia

This approach to generate candidate terms for a given query uses Wikipedia as an external knowledge source. Wikipedia has millions of concepts, and it is not easy to find suitable terms which can serve as expansion of a specific query that is,

$$\mathrm{Qt} = \{t_1, t_2, \ldots, t_{\mathrm{m}}\} \tag{8}$$

Where t_i is a term of initial query and *m* is the number of terms. We have several steps for generating the candidates:

First, the full-content of Wikipedia is preprocessed and only text is used, illegal characters and stop words are filtered, terms are stemmed using the Porter stemmer for English texts [23] and converted into lowercase. Then the full-text is indexed, so each concept is mapped to all the terms that are important for it. After scanning this index of Wikipedia concepts, we represent the query as interpretation vector in the space of concepts:

$$\mathrm{Qc} = \{c_1, c_2, \ldots, c_{\mathrm{n}}\} \tag{9}$$

Where *n* is the size of the query concept vector. This is the process of Explicit Semantic Analysis [6]. This representation serves in Boosting technique inspired from Lucene [24] which allows us to control the relevance of a query by boosting its concepts. The boosting factor for each concept is represented by its score in this vector (9). The higher the boost factor, the more relevant the concept will be. Then a relevance between the query concepts vector and index of Wikipedia is computed to perform term retrieval. This relevance is mainly determined by a scoring formula based on TF.IDF implemented in Lucene [24], it actually reflects the similarity between the query and each Wikipedia term. Formally, this relevance is defined as:

$$R(Qc,t) = coord(Qc,t) \times norm(Qc) \times \sum_{c_k \in Q_c} (tf_t^{c_k} \times (idf_t^{w})^2 \times c_{boost}) \tag{10}$$

Where *t* is a Wikipedia term, tf_t^{ck} is the term frequency of *t* in the concept c_k , *W* is the whole set of Wikipedia concepts, and idf_t^w is the *idf* value of *t* over *W*, c_{boost} is the specified boost for concept c_k, *coord (Qc,t)* is the number of concepts in both *Qc* and *t* divided by the number of concepts in *Qc*, *norm(Qc)* is a normalizing factor used to make scores between queries comparable.

3.2 Building the query graph

In the first phase of the approach the irrelevant terms have been filtered out with the relevance threshold, but we don't have any relatedness between the retained terms and we risk to keep candidates with a high relevance threshold and less quality. Thus, in this second stage we need to take into account the influence among Wikipedia terms to find the most important ones. In Wikipedia, terms are semantically related. Logically when a term t_1 is selected for expansion of the query *Qt*, the term t_2 which is quite close to t_1, should also be selected. Therefore, we construct a query graph based on the ranked terms, in which each node represents a term, and each edge measures the relatedness between the two corresponding terms.

Fig. 2 presents an illustration of a query graph. The relatedness *R* between two terms or query and term is calculated using Eq. (10).

We have main steps for building the query graph. First, a set of terms generated in the previous step for the given query are used as initial vertices t_j and edges measures the relatedness $R(Q_c,t_j)$ between query and terms. This iterative process of generating related terms is repeated for each new vertex t_j , new vertices t_i are branched with relatedness $R(t_j,t_i)$ while the number of iterations is not achieved. As result we have an oriented query graph, which presents the most related terms to the query, but the number of candidates is still too large for query expansion. For this reason, the next phase consists in ranking the candidate terms in order to ensure that the queries will contain adequate terms.

3.3 Candidate terms filtering

For avoiding the inclusion of non-adequate terms in the extended queries, we propose a new relatedness measures for estimating the final relatedness between a candidate term and a given query, these measures combine the association rules confidence, a new

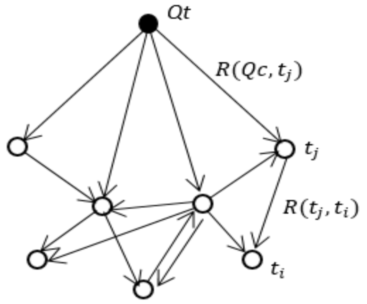

Figure. 2 An illustration of query graph-based term

measure deducted from the query graph and ESA. Each of them has positive impact on the selection of the best expansion terms, and their combination promotes a considerable improvement.

3.3.1. Graph similarity

An important form of information is the links among terms in the query graph. These generated links represent relatedness between the candidate terms. Intuitively, if a term is both linked to many other terms and directly linked to the query, it is much likely to have a strong semantic relation with the query. For instance, “HIV” and “Prevention” are candidate terms of query “AIDS Treatment”, the fact that the first one is more linked to other terms provides evidence that “HIV” is more related to the query than “prevention” (see Fig. 3).

The semantic relatedness between a query and candidate term is determined by two factors: (1) the term is directly linked to the query in the graph; (2) the number of other terms which have direct link to this term. Then the relatedness between query *Qt* and term *t* deducted from the query graph is described in the following formula:

$$QG\_score(Qt,t) = \beta \times R(Qc,t) + (1-\beta) \times \frac{\sum_{t_i \in T} R(t,t_i)}{\max_{t_i \in T} R(t,t_i) \times n_{total}} \quad (11)$$

Where *R(Qc, t)* is a similarity between the query concepts vector *Qc* and term *t, R(t,t_i)* is similarity between two terms t and t_i which have already been mentioned in Eq. (10) above ,*T* is a set of query graph terms directly related to *t*, n_{total} is the total number of terms in *T*, and β is a weighting parameter $\in$ [0,1] to control the influence of query relevance and other terms.

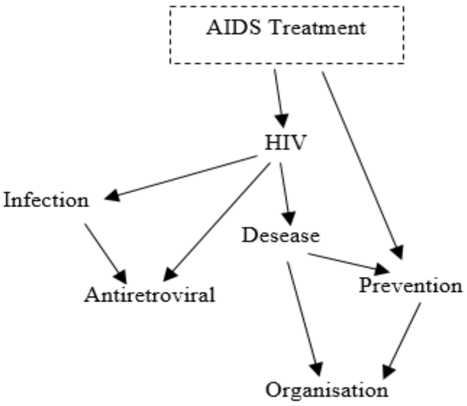

Figure. 3 An illustration of query graph of ”AIDS Treatment”

3.3.2. Association rules similarity

The idea is to use the association rules mining technique to discover strength correlations between query and candidate terms in documents collection. The process of generating association rules for a given query is performed as in the following steps: First, we select a set of unstructured full texts related to the original query, from documents collection using TF-IDF. Second, the selected texts are tokenized and only text is used, illegal characters and stop words are filtered, sentences are identified, terms are stemmed using Porter stemmer [23] for English texts and converted into lowercase, and the preprocessed documents are saved. As second step, we construct the transactional dataset by considering each keyword as item, each sentence as transaction and the document in which the sentence occurs as transaction elements. After we import a transactional dataset and we apply the referenced algorithm in [18]. It is based on ELECTRE [25] method which is able to select the most interesting association rules generated using Apriori [26] by considering a new outranking relation. The main advantage of this method is that it is not hindered by the abundance of measures and it evaluates the association rules using a set of criteria, not only one.

We exploit the association rules extracted for controlling if the candidate term is adequate for query expansion. The maximum of the confidence of any association rule that contains at least one of the query terms and the candidate term is retained as score for this last one, then this score is defined as:

$$Conf_{\max}(R,Qt,t) = \max_{t_{qt} \in Q_t, R_j \in R} Conf(R_j(t_{qt},t)) \quad (12)$$

Where R_j is an association rule from *R*, t_{qt} is term from query *Qt* and *t* is the candidate term.

3.3.3. Combining relatedness measures

The relatedness measure between the original query and each candidate term is inferred from two factors presented above. This measure is defined as:

$$\text{Score(Qt,t)=}\begin{cases} \alpha \times QG\_score(Qt,t) + (1-\alpha) \times Conf_{\max}(R,Qt,t) \\ \qquad Or \\ \alpha \times QG\_score(Qt,t) + (1-\alpha) \times ESA(Qt,t) \end{cases} \quad (13)$$

Where *QG_score(Qt,t)* is the deducted query graph score in Eq. (11), *Conf_{max}(R,Qt,t)* is the association rules score in Eq. (12) and *ESA(Qt,t)* is the score mentioned in Eq. (4). α is a weighting parameter ∈ [0,1].

Once the semantic relatedness between a query and its candidate terms is calculated, we select the most related ones and we add them to the original query.

4. Experiments and analysis

In this last section, we present experimental studies to test the effectiveness of the proposed approach. Before going to the details, we first present the evaluation metrics and test collection on which our runs were conducted.

4.1 Test collection and runs

The collection of TREC AP8889 issued from the Text Retrieval Conference (TREC) project that is a set of news articles written in English and published by Associated Press over a period of 2 years (1988-1989) has been chosen to apply our proposed approach. In our work, this collection is indexed using Lucene [24], which is a free and open source information retrieval library, completely written in java. The same library is then used for retrieving the top 1,000 documents, for each query from topics using the Okapi BM25 model [27].

The collection chosen contains 164597 documents with 150 topics from which the queries are extracted and a relevance judgments file done by domain experts. Only titles of the TREC topics are used as queries for simulating search scenarios where users tend to submit short queries. Comparing the responses of a system according to a query with a relevance judgement allows us to evaluate the following metric:

*Precision: measures the proportion of relevant documents among all documents retrieved by the system.

We expand each query in TREC collection with the expansion terms retrieved by our approach. The expanded queries are answered by information retrieval system based on Lucene, and the generated responses are evaluated. Our baseline is the 0-Expansion method, which means the original queries are interrogated without any expansion.

Critical steps in our approach are Query Graph, association rules technique and the process of ESA. In order to test the effects of these steps in filtering phase, we exclude them from our system, and we expand each query with the top terms extracted from Wikipedia using one or two techniques of them, and we obtain the following runs:

*0-Expansion: the baseline

*0-filtering: Query expansion without any filtering phase.

*QG: Query expansion based on graph similarity in filtering phase.

*ESA: Query expansion based on ESA similarity in filtering phase.

*AR: Query expansion based on association rules similarity in filtering phase.

*QG-ESA: Query expansion based on graph and ESA scores in filtering phase.

*QG-AR: Query expansion based on graph and association rules similarities in filtering phase.

In all these runs the expansion terms are extracted from the Query Graph except the first one. The parameters are set in the following ways:

*The query-term relevance threshold for candidate selection from Wikipedia is empirically set to 0.6.

*Number of iterations for query graph building is set to 2.

*The parameters for association rules algorithm are determined by taking minimal values for not excluding any important rule: minSupport = minConfidence= minLift= minGain= minJacard= minCos=0.1.

*For the controlling factor α and β, we experiment on a wide range of values and choose the best one in terms of the evaluation metrics (α=0.5, β=0.7).

*We select the number of expansion terms from {2,3,4,5}.

In order to evaluate the performance of the proposed approaches, we must compare the obtained experimental results with some recent research contributions in the domain of query expansion based on Wikipedia [15 - 17]. However, an inconsistency in results is detected even using the same approaches and test corpus. These inconsistencies can be explained by the fact that the running of indexing and retrieving phases used a large variety of optimization parameters such as stop word elimination, stemming algorithms, ranking methods, etc. Consequently, for comparing our approaches, we have used the same search engine Lucene [24] with the same parameters values and test corpus TREC AP8889 in implementing some existing methods recently proposed which are:

*QE-Wiki-Hyperlinks: QE Powered by Wikipedia Hyperlinks [17].

*QE-Wikimantic: Query expansion via Wikimantic [15].

Table 2. Performance comparisons using MAP of our runs with respect to the baseline and the three existing algorithms

Run	MAP	Map-Gain
QG-AR	0,3667	39%
QG-ESA	0,3606	37%
Feedback-QE-Retrieval [16]	0.3518	34%
QE-Wiki-Hyperlinks [17]	0.3408	29%
AR	0,333	26%
QG	0,3178	21%
QE-Wikimantic [15]	0.299	13%
ESA	0.2975	13%
0-filtering	0,2337	-11%
0-Expansion	0,2635	-

*Feedback-QE-Retrieval: feedback-query-expansion based candidates generation [16].

4.2 Results and discussion

The results achieved by the proposed methods, the baseline and the three existing algorithms are summarized in Table 2 and Fig. 4.

The results in Table 2 show that our systems QG-AR and QG-ESA using the proposed relatedness measures that combines two similarities lead to a significant improvement compared to the baseline and other runs in terms of Mean Average Precision (MAP). For instance, when evaluated on MAP, QG-AR and QG- ESA have approximately 39% and 37% improvement over 0-Expansion respectively, while the improvements achieved by Feedback-QE-Retrieval and QE-Wiki-Hyperlinks are 34% and 29% respectively. In the case of AR, QG, ESA and QE-Wikimantic, some improvements have been presented, but less than 26%. For 0-filtering run, it did not perform well, despite of extracting terms from Wikipedia. It is due to noise coming from the presence of unrelated terms to the query and absence of any filtering phase.

Fig. 4 shows the precision when x documents are retrieved (P@X) and X is set to 5, 10, 15 and 20 respectively. It can be seen that using graph terms of Wikipedia for query expansion without excluding the filtering step leads to the improvement of the retrieval effectiveness in comparison with the baseline. Filtering step is the main phase in query expansion which ensured the extended queries containing adequate terms.

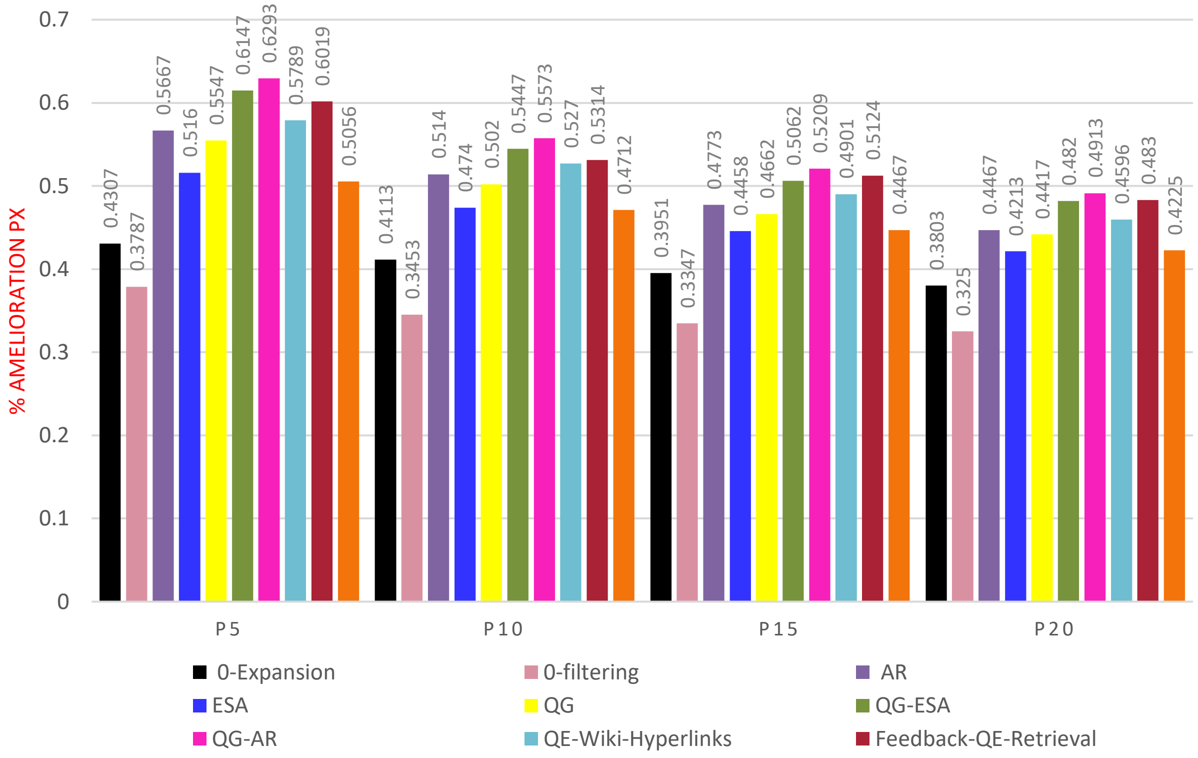

Figure. 4 Improvement percentage in P@X

For example, the precisions for the top five and ten retrieved documents using QG-AR achieve the highest scores 0.6293 and 0.5573, respectively. While the baseline brings only 0,4307 (+46%) and 0.4113 (+35%).For the QG-ESA 0.6147(+42) and 0.5447(+32) and for Feedback-QE-Retrieval 0.6019 (+40%) and 0.5314(+29%). These experiments have the advantage to rank the relevant documents according to queries in the top of results.

The good performance obtained is explained by the use of the proposed approaches which are based on high quality terms extracting from query graph generated from Wikipedia articles with respect to the relation between terms. When ESA, AR, QG aim to filter the expansion terms by considering one of the following factors:
*the expansion term is semantically relevant with the query (ESA).
*the expansion term is strongly correlated with the query terms in an unstructured texts contents (AR).
*the expansion term is quite close to other terms (QG).

The proposed QG-AR and QG-ESA aim to cover two factors for capturing the most important terms by combining two similarities in filtering step which allows to expand the queries with a relevant term. For Query expansion via Wikimantic approach proposed by [15], the first retrieval to construct a collection of relevant documents for creating candidate terms is the most important phase, then weights are applied to the concepts. The collection of documents may concern certain domains. In spite of the precision in the process of selecting relevant terms, irrelevant ones can be added to the user query. While QE-Wiki-Hyperlinks, described in this paper [17], adds a variety of quality expansion terms in some contexts to the original query, due to coming directly from Wikipedia. This is considered as advantage in some way. However, the terms may be irrelevant. Feedback-QE-Retrieval discussed in [16], using Wikipedia terms based on co-occurrence mentions. The data used includes article titles, article content and article redirections for building the collection of enhancement which serves as feedback for finding expansion terms.

The overall performance of the proposed approaches is promising in terms of precision in comparison with existing techniques. This performance proves that a combination between Wikipedia graph, association rules based on multi-criteria optimization and ESA can improve the retrieval effectiveness by expanding the queries semantically.

5. Conclusion and future work

In this paper, query expansion approach which expands queries with Wikipedia terms extracted by the query graph model is proposed. These terms are filtered using a new relatedness measure that combines the strength of the semantic similarity presented by ESA, the closeness between terms in the query graph and association rules technique based on multi-criteria optimization. The experimental study was conducted on TREC AP8889. The results are very promising, and our approach outperforms the baseline method significantly and others approaches based on Wikipedia. In terms of Mean Average Precision (MAP) the proposed approaches have approximately 39% and 37% improvement over the baseline, while the improvements achieved by the comparison methods recently proposed don’t exceed 34%. These results approve that our query expansion approach that combines Wikipedia and text mining technique is an effective way to improve the performance of information retrieval systems. For future works, we propose to use an ontology as an external data sources and other forms of information for query expansion. We will also use other text mining algorithms.

References

[1] G. Cao, J. Y. Nie, J. Gao, and S. Robertson, “Selecting Good Expansion Terms for Pseudo-Relevance Feedback”, In: *Proc. of the 31st ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.243-250, 2008.

[2] T. Tao and C. Zhai, “Regularized Estimation of Mixture Models for Robust Pseudo-Relevance Feedback”, In: *Proc. of the 29th ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162-169, 2006.

[3] Y. Xu, G. J. F. Jones and B. Wang, “Query Dependent Pseudo-Relevance Feedback based on Wikipedia”, In: *Proc. of the 32nd ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59-66, 2009.

[4] G. Salton and C. Buckley, “Improving retrieval performance by relevance feedback”, *Journal of the American Society for Information Science,* Vol.14, pp.288–297, 1990.

[5] M. A. Zingla, L. Chiraz, and Y. Slimani, “Short Query Expansion for Microblog Retrieval”, *Procedia Computer Science*, Vol. 96, pp. 225–234, 2016.

[6] E. Gabrilovich and S. Markovitch, “Computing Semantic Relatedness using Wikipedia-based

Explicit Semantic Analysis", *IJcAI*, Vol.7, pp. 1606–1611, 2007.

[7] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases", *ACM sigmod record*, Vol. 22, No. 2, pp. 207–216, 1993.

[8] C. Lv, R. Qiang, F. Fan, and J. Yang, "Knowledge-based query expansion in real-time microblog search", In: *Proc. of Asia Information Retrieval Symposium*, pp. 43-55,2015.

[9] R. Li, L. Hao, X. Zhao , P. Zhang , D. Song, and Y. Hou, "A query expansion approach using entity distribution based on markov random fields", In: *Proc. of Asia Information Retrieval Symposium*, pp.387–393, 2015.

[10] W. C. Brandao , "Exploiting entities for query expansion". In: *Proc. of ACM SIGIR Forum*, Vol.48, No. 1, pp. 43-43, 2014.

[11] B. El Ghali, A. El Qadi, M. Ouadou and D. Aboutajdine, "Context-based query expansion method for short queries using latent semantic analyses". In: *Proc. of International Conference on Networked Systems*, pp. 468-473, 2015.

[12] B. El Ghali and A. El Qadi, "Context-aware query expansion method using language models and latent semantic analyses", *Knowledge and Information Systems*, Vol.50, No.3, pp. 751-762, 2016.

[13] R. Anand and A. Kotov, "An empirical comparison of statistical term association graphs with dbpedia and conceptnet for query expansion", In: *Proc. of the 7th Forum for Information Retrieval Evaluation*, pp.27–30, 2015.

[14] A. Jain, K. Mittal and DK. Tayal, "Automatically incorporating context meaning for query expansion using graph connectivity measures", *Progress in Artificial Intelligence*, Vol.2, No.2-3, pp.129–139 ,2014.

[15] C. Boston, H. Fang, S. Carberry, H. Wu, and X. Liu, "Wikimantic: Toward effective disambiguation and expansion of queries", *Data & Knowledge Engineering*, Vol. 90, pp.22 – 37, 2014.

[16] G. Zhao, J. Wu, D. Wang, and T. Li, "Entity disambiguation to Wikipedia using collective ranking", *Information Processing & Management*, Vol.52, No.6, pp.1247 – 1257, 2016.

[17] C. Bruce, X. Gao, P. Andreae, and S. Jabeen, "Query Expansion Powered by Wikipedia Hyperlinks", In: *Proc. of Australasian Joint Conference on Artificial Intelligence*, pp.421–432, 2012.

[18] A. Dahbi, S. Jabri, Y. Ballouki, and T. Gadi, "A new method to select the interesting association rules with multiple criteria", *International Journal of Intelligent Engineering and Systems*, Vol.10, No.5, pp.191-200, 2017.

[19] H. P. Luhn, "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development*, Vol.2, No.2, pp.159-165, 1958.

[20] G. Salton, "Introduction to Modern Information Retrieval", *McGraw-Hill*, 1983.

[21] CJ. Van Rijsbergen, "Information retrieval", *Butterworths*, 1979.

[22] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", *Information processing & management*, Vol.24, No.5, pp. 513-523, 1988.

[23] M. Porter PorterStemmer (java version) [Software], In https://tartarus.org/ martin/ PorterStemmer/index-old.html, 1980.

[24] Lucene, http:lucene.apache.org/core, last accessed 2017/05/01.

[25] J. Dyer, "Multiple Criteria Decision Analysis: State of the Art Surveys", *Springer-Verlag*, Vol. 78, New York, pp. 265-292, 2005.

[26] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In: Proc. *of the 20th Int. Conf. Very Large Data Bases VLDB*, pp. 487-499, 1994.

[27] S. E. Robertson, S. Walker, M. Beaulieu, and P. Willett, "Okapi at trec- 7: Automatic ad hoc, filtering, vlc and interactive track", *Nist Special Publication SP*, No.500, pp.253–264, 1999.