Inf Retrieval J (2018) 21:337–367
https://doi.org/10.1007/s10791-017-9326-6

Hybrid query expansion model for text and microblog information retrieval

Meriem Amina Zingla[1] · Chiraz Latiri[2] · Philippe Mulhem[3] · Catherine Berrut[3] · Yahya Slimani[2]

Received: 10 July 2016 / Accepted: 21 December 2017 / Published online: 3 February 2018

Abstract Query expansion (QE) is an important process in information retrieval applications that improves the user query and helps in retrieving relevant results. In this paper, we introduce a hybrid query expansion model (HQE) that investigates how external resources can be combined to association rules mining and used to enhance expansion terms generation and selection. The HQE model can be processed in different configurations, starting from methods based on association rules and combining it with external knowledge. The HQE model handles the two main phases of a QE process, namely: the candidate terms generation phase and the selection phase. We propose for the first phase, statistical, semantic and conceptual methods to generate new related terms for a given query. For the second phase, we introduce a similarity measure, ESAC, based on the Explicit Semantic Analysis that computes the relatedness between a query and the set of candidate terms. The performance of the proposed HQE model is evaluated within two experimental validations. The first one addresses the tweet search task proposed by TREC Microblog Track 2011 and an ad-hoc IR task related to the hard topics of the TREC Robust 2004. The second experimental validation concerns the tweet contextualization task

✉ Meriem Amina Zingla
meriem.zingla@etudiant-fst.utm.tn

Chiraz Latiri
chiraz.latiri@gnet.tn

Philippe Mulhem
philippe.mulhem@imag.fr

Catherine Berrut
catherine.berrut@imag.fr

Yahya Slimani
yahya.slimani@gmail.com

[1] Faculty of Sciences of Tunis, Tunis EL Manar University, Campus Universitaire Farhat Hached, B.P. 94, 1068 Tunis, Tunisia

[2] Higher Institute of Multimedia Arts of Manouba, Manouba University, Manouba, Tunisia

[3] MRIM Group, LIG laboratory, Grenoble Alpes University, Grenoble, France

organized by INEX 2014. Global results highlighted the effectiveness of our HQE model and of association rules mining for QE combined with external resources.

Keywords Information retrieval · Query expansion · Tweets search · Explicit Semantic Analysis · Tweet contextualization · WIKIPEDIA · DBPEDIA · Association rules · Ad-hoc IR task

1 Introduction and motivations

In text information retrieval (IR), query expansion (QE) refers to the techniques and algorithms that reformulate the original query by adding new terms into the query, in order to improve the retrieval effectiveness. Many query expansion techniques were developed in the past decades. In this respect, an interesting survey on QE is given in Carpineto and Romano (2012).

In the literature, query expansion approaches are mainly classified as *global* (Xu et al. 1996), *local* (Buckley et al. 1994) and *external* approaches. Roughly speaking, global approaches expand the query by adding new query terms that are statistically related with the initial query terms. On the other hand, local approaches use retrieved documents produced by the initial query, and mostly use pseudo-relevance feedback (PRF) (Xu et al. 1996). Whereas, the external approaches rely on external resources such as encyclopedic knowledge extracted from WIKIPEDIA (Li et al. 2007), or conceptual ones which are derived from ontologies (Bhogal et al. 2007). In addition, hybrid QE approaches, that rely on the combination of two (or more) QE methods, are also possible. Some examples are those in Han and Chen (2009), Ko et al. (2008). These approaches will be described in the next section.

In this paper, while considering textual and microblog collections, our contributions address tweet search, ad-hoc IR and the tweets contextualization tracks. These tracks are the result of the explosive growth of textual resources on the web, especially in microblogs. In fact, microblog retrieval has drawn tramendous attention in recent years. Therefore, TREC introduced a track for ad-hoc microblog retrieval in 2011 (Ounis et al. 2011) where large tweet collections and annotations for various queries were released. In this respect, different approaches were investigated for microblog retrieval to overcome the special nature of microblog messages, e.g., short, noisy and time-sensitive characters of microblog posts. However, one of the main challenges in microblog retrieval is term mismatch due to short queries. In the recent literature, the term mismatch problem in microblog posts is tackled through various techniques (Meij et al. 2012; Jabeur et al. 2012). Among them, we are interested in those using query expansion (Lau et al. 2011; Massoudi et al. 2011; Bandyopadhyay et al. 2012; Shekarpour et al. 2013; Lv et al. 2015). Our proposed model for expanding queries is divided into two main phases, namely *terms generation* and *the terms selection*. Our work focuses on both the text retrieval task and the microblog search results by using, on one hand, the implicit knowledge provided by advanced text mining methods, especially association rules (Agrawal and Skirant 1994); and knowledge extracted from external resources such as WIKIPEDIA and DBPEDIA[1] (Aggarwal and Buitelaar 2012), on the other hand. Moreover, we tackle the issue of expansion terms selection with respect to the semantic relatedness between original query terms and candidate terms (Luo et al. 2012; Klyuev and Haralambous 2011; Bouchoucha et al. 2014).

[1] http://dbpedia.org/.

Thereby, and in order to enhance expansion term generation, we rely on the use of association rules (Agrawal and Skirant 1994) between terms which consists in extracting relations between terms based on a global analysis of a document collection. Those association rules convey statistical relations between terms that are used in an automatic query expansion process. It is also interesting to note that the QE approaches based on association rules do not require a priori knowledge or a complicated linguistic process. They are based on an automatic process without any external or human intervention nor any external knowledge resources (thesaurus, ontology, etc.). The use of association rules highlighted their efficiency in the IR field in previous studies, as in Martín-Bautista et al. (2004), Song et al. (2007), Latiri et al. (2012), Wei et al. (2000), Liu et al. (2013), Belalem et al. (2016). In fact, the extraction of association rules between terms is performed in two steps: the first step consists in extracting termsets, i.e., sets of terms, in a document collection that can be reasonably represented as a family of subsets of terms from a global set. A document collection can then be seen as a family of termsets drawn from a global set of index terms. Whereas, the second step consists in generating the association rules. An association rule is a relation $T_1 \Rightarrow T_2$, where T_1 and T_2 are two termsets. The advantage of the insight gained through association rules is in the contextual nature of the discovered inter-term correlations. Thus, the confidence of an association rule approximates the probability of having the terms of the conclusion in a document, given that those of the premise are already there.

In this paper, we investigate how external resources and PRF can be combined to association rules mining and used to enhance expansion terms generation and selection. This leads to a hybrid approach to handle query expansion, denoted HQE in the remainder of the paper, that proposes an efficient synergy between local, global and external QE methods. We propose three approaches for incorporating additional knowledge when generating expansion terms, namely: (i) a *statistical approach* which relies on association rules mining to discover strength correlations between terms, handled as a local method (PRF) combined with a global method. So, if the original query terms are included in the premises of mined rules, they will be thus expanded using the set of terms contained in the conclusion parts of selected rules; (ii) a *semantic approach* which consists in exploring WIKIPEDIA[2] articles, especially, the articles definitions parts, and extracting information from these latter to expand the original query; and, (iii) a *conceptual approach* based on the DBPEDIA ontology. This approach consists in accessing the DBPEDIA data set on the Web and extracting related information for a given query. The two last approaches, i.e., semantic and conceptual are part of external QE methods. Furthermore, the proposed HQE model can be applied in different configurations, starting from the statistical method based on the association rules and combining it with the semantic and conceptual knowledge. The driving idea behind combining these methods is to obtain performance results much better than that of the individual best results. This is achieved by combining several independent query expansion results and choosing the best results that outperform the baseline.

In addition to our efficient term generation, we also propose to enhance the selection of good expansion terms, by introducing a new semantic relatedness measure, named ESAC.

This measure combines the WIKIPEDIA-based Explicit Semantic Analysis (ESA) measure (Gabrilovich and Markovitch 2007) and the confidence metric of association rules (Agrawal and Skirant 1994). It allows to estimate a semantic relatedness score between the query and its relevant terms extracted using association rules. We note that the proposed

[2] https://www.wikipedia.org/.

measure ESAC considers both encyclopedic and correlation knowledge about terms. This advantage of ESAC is a key factor to find precise terms for automatic query expansion.

We validate the proposed HQE model on two kinds of evaluations. The first experiments are devoted to IR tasks in the case of *difficult cases* for which potential mismatches between queries and documents, namely TREC 2011 microblog search and TREC Robust 2004 ad-hoc tracks. We thoroughly evaluate to what degree our proposals aid retrieval effectiveness. The second experimental validation is dedicated to the tweet contextualisation task with INEX 2013 and 2014, aiming at providing, for a given tweet, a context from WIKIPEDIA articles, in a way that makes it clear for a reader. In this case, our HQE model is able to extend, properly, the original query tweets in a way to retrieve relevant and diverse WIKIPEDIA documents that lead to a higher context quality.

The remainder of the paper is organized as follows: Sect. 2 discusses related works on query expansion for information retrieval. Section 3 introduces some basic definitions related to our work. Then, in Sect. 4, a detailed description of our HQE model for query expansion is presented. Section 5 is devoted to experimental validation within two IR ad-hoc text and microblog tasks. In Sect. 6, we describe the embedding of the proposed query expansion model for INEX tweet contextualization task. The "Conclusion" section wraps up the article and outlines future works.

2 Related work

In this section, we discuss the query expansion approaches, and elicit our HQE model based on statistical, semantic and conceptual methods for generating candidate terms expansion. Hence, these latter can be either extracted from external resources such as WIKIPEDIA, DBPEDIA, etc., known as *external resources based QE* (Al-Shboul and Myaeng 2014) or from the documents themselves; based on their links with the initial query terms, named *document based QE*. In the literature, document based QE approaches contain two major classes: *global* approaches and *local* approaches (Carpineto and Romano 2012). Here, we will mention the efforts on both of them.

Local QE methods use retrieved documents produced by the initial query. It mainly refers to relevance feedback and pseudo relevance feedback (Buckley et al. 1994) approaches to reformulate the query. These methods use top-ranked documents retrieved by the initial original query. However, the top retrieved documents may not always provide good terms for expansion, particularly for difficult or short queries with few relevant documents in the collection which do not share relevant terms. These methods lead to topic drift and negatively impact the results (Macdonald and Ounis 2007).

Authors in Cao et al. (2008) re-examined the assumption which provides that PRF assumes that most frequent terms in the pseudo-feedback documents are useful for the retrieval does not hold in reality. In Chen and Lu (2010), authors showed that relevant expansion terms can not be distinguished from bad ones merely on their distributions in the feedback documents and in the whole collection. They proposed to integrate a term classification process to predict the usefulness of expansion terms. Recently, in Colace et al. (2015) authors have demonstrated the effectiveness of a new expansion method that extracts weighted word pairs from relevant or pseudo-relevant documents. They have also applied learning to rank methods to select useful terms from a set of candidate expansion terms within a PRF framework. Their obtained results demonstrated that the QE method based on their new structure outperforms the baseline (Colace et al. 2015). Moreover, to

take advantage of the word embeddings representation, in Almasri et al. (2016), authors explored the use of the relationships extracted from deep learning vectors for QE. They showed that word embeddings are a promising source for query expansion by comparing it with PRF and the expansion method using mutual information.

Global QE methods unlike local QE, in global methods, candidate terms come from the entire document collection rather than just (pseudo-) relevant documents. In Xu et al. (1996), authors proved that using global analysis techniques produces results that are both more effective and more predictable than simple local feedback. Such QE approaches are generally based on extraction of relationships between terms among the whole document collection and based on their co-occurrences where the window size used is a document.

In Järvelin et al. (2001), authors developed a deductive data model for concept-based query expansion. It is based on three abstraction levels: the conceptual, linguistic and string levels. Concepts and relationships among them are represented at the conceptual level. The linguistic level gives natural language expressions for concepts. Each expression has one or more matching patterns at the string level. In Gong et al. (2006), the authors used WordNet and the Term Semantic Network (TSN) for developing word co-occurrence-based Thesauri. The TSN was used as a filter and provided a supplement for WordNet. However, it was noticed that the Thesauri construction strategy was complex and tedious.

In addition to the global approach based on Thesaurus construction, we focus on association rules mining which targets to retrieve correlated patterns (Agrawal and Skirant 1994) from the documents collection. An association rule binds two sets of terms namely *a premise* and *a conclusion*. This means that the conclusion occurs whenever the premise is observed in the set of documents. To each association rule, a confidence value is assigned to measure the likelihood of the association. It has been proven, in the literature, that the use of such dependencies for QE could significantly increase the retrieval effectiveness (Wei et al. 2000). (i.e.) Association rules reflect implicit and strong correlations between terms. Using these correlations for expanding queries allows to enrich the query presentation by adding a set of related terms and consequently improve retrieval performance by matching additional documents. Hence, the authors in Tangpong and Rungsawang (2000) performed a small improvement when using the APRIORI algorithm (Agrawal and Skirant 1994) with a high confidence threshold (more than 50%) that generated a small amount of association rules. Using a lower confidence threshold (10%), authors performed better results (Tangpong and Rungsawang 2000). In Haddad et al. (2000), authors proposed the same approach performing improvement when using the APRIORI algorithm to extract association rules. The best improvements were performed with low confidence values. The approach in Martín-Bautista et al. (2004) has refined the query based on association rules. Given an initial set of documents retrieved from the web, text transactions are constructed and association rules are formulated. These rules are used by the user to add additional terms to the query for improving the precision of the retrieval. A more adapted mining algorithm to text that avoids redundancy in mined association rules is proposed in Latiri et al. (2012). Non redundant association rules between terms are used to expand the user query considering all terms that appear in the conclusions of these rules whose premise is contained by the original query. Experimental evaluation of this approach shows an improvement of the IR task. Closer to our work, in Song et al. (2007), the authors proposed a novel semantic query expansion technique that combines association rules with ontologies and Natural Language Processing techniques. This technique uses the explicit semantics as well as other linguistic properties of unstructured text corpus. It incorporates contextual properties of important terms discovered by association rules, and ontology entries are added to the query by disambiguating word senses.

External QE methods external QE approaches involve methods that generate expansion terms from other resources besides the target corpus. Many approaches used the WIKIPEDIA Corpus for query expansion as it is the biggest encyclopedia and is freely available on the web. Although it has been manually developed, its contents are well structured and growing rapidly with a wide variety of topics, which makes it a good knowledge source for query expansion. In Li et al. (2007), authors used the WIKIPEDIA corpus for the query expansion by utilizing the category assignments of the articles. The initial query was used against the WIKIPEDIA collection and each category was assigned a weight that was proportional to the number of top-ranked articles assigned to it. The articles were re-ranked based on the sum of the weights of the categories to which they belonged. For short query expansion, authors in Almasri et al. (2013), proposed a semantic approach that expands short queries by semantically related terms extracted from WIKIPEDIA. Recently, authors in Gan and Hong (2015) proposed a new approach to extract more term relationships from Markov network for query expansion, where term relationship extracted from WIKIPEDIA corpus is superimposed to the basic Markov network pre-built using a single local corpus.

Nevertheless, the aforementioned methods rely on counting word co-occurrences in the documents to select expansion terms knowing that they are not always a good indicator for relevance, whereas some are background words of the whole collection. In order to select good expansion terms, Explicit Semantic Analysis (ESA) is adopted in some contributions such as Luo et al. (2012) where authors used ESA to estimate two kinds of relevance weight. One is the relevance weight between query and its relevant word extracted from the top-ranked documents in initial retrieval results. The other is the relevance weight between each query word and its relevant words extracted from the snapshot of Google search result when that query word is used as search keyword. The estimated relevance weights are used to select good expansion words for second retrieval. Klyuev and Haralambous (2011) investigated the efficiency of the proposed EWC semantic relatedness measure in an ad-hoc retrieval task. This measure combines the WIKIPEDIA-based Explicit Semantic Analysis measure, the WordNet path measure and the mixed collocation index. Conducted experiments demonstrated promising results.

Furthermore, Hybrid approaches have achieved a promising results in tackling query expansion issues. Authors, in Ko et al. (2008), use a statistical query expansion technique along with pseudo relevance feedback and query summarization techniques. They try to generate an effective snippet at the beginning as compared to other traditional methods. Authors in Selvaretnam et al. (2013) use linguistic and statistical techniques for query structure classification for the application to query expansion. Authors in Han and Chen (2009) propose a hybrid method for query expansion (HQE). HQE method is a combination of ontology-based collaborative filtering and neural networks. The ontology-based collaborative filtering is used to analyze the semantic relation, while radial basis function networks are used to retrieve the most relevant web documents. Their method can enhance the precision and also user can provide less query information at the beginning as compared to other traditional methods.

In addition, since the proliferation of microbloging platforms, dealing with microblogs has become increasingly important, and as these microblogs messages are short and are, to some extent, ambiguous, QE has been widely used in microblog retrieval, such as tweet retrieval (Lv et al. 2015). Bandyopadhyay et al. (2012) used external corpora as a source for query expansion terms. They relied on the Google Search API (GSA) to retrieve pages from the Web, and expanded the queries employing their titles. Lau et al. (2011) proposed a twitter retrieval framework that focuses on topical features, combined with query expansion using PRF to improve microblog retrieval results. Massoudi et al. (2011)

developed a language modeling approach tailored to microblogging characteristics, where redundancy-based IR approaches cannot be used in a straightforward manner. They enhanced this model with two groups of quality indicators: textual and microblog specific. Additionally, they proposed a dynamic query expansion model for microblog post retrieval.

QE is also used to expand microblog posts. A popular task is INEX Tweet contextualization task (Bellot et al. 2016) which addresses the problem of tweet enrichment in order to generate its context and make it more understandable. In Morchid et al. (2013), authors used Latent Dirichlet Analysis (LDA) to obtain a representation of the tweet in a thematic space. This representation allows to find a set of latent topics covered by the tweet. This approach gives good results for the tweet contextualization task. In Zingla et al. (2014), association rules between terms are used to extend the tweet. Authors projected the terms of the tweet on the rules' premises and added the conclusions to the original tweets. Obtained results highlighted an interesting improvement within the tweet contextualization task.

In this paper, we propose to revisit these QE approaches proposing a hybrid QE model, where the generation of the expansion terms relies on local, global and external QE methods. The driving idea behind our proposal is to enhance QE results since statistical, semantic and conceptual methods are combined to generate new related terms for a given query. These methods use, respectively, correlation knowledge represented by association rules between terms, semantic knowledge from WIKIPEDIA and conceptual knowledge extracted from the DBPEDIA ontology. Furthermore, using a new relatedness measure ESAC, the proposed HQE model leads to different configurations that we validate on two ad-hoc IR tasks and one contextualization tweet task.

It is worth noting that this paper is a large extension of Zingla et al. (2016) as it involves a more detailed formalization of the HQE model and more developed experiments on different IR tasks.

In the next section, we introduce the basic definitions related to our proposed query expansion model.

3 Basic definitions

As aforementioned, query expansion is a technique utilized within information retrieval to solve word mismatch between queries and documents. Expansion words are usually selected by counting word co-occurrences in the documents. However, word co-occurrences are not always a good indicator for relevance, whereas some are background words of the whole collection. To elevate this issue, we introduce a QE model with twofold improvements with respect to the candidate terms for expansion, namely: (1) we rely on association rules between terms to derive efficient candidate terms (Latiri et al. 2012); and (2) in order to select good expansion terms, explicit semantic analysis (ESA) is adopted in our model to estimate a semantic relatedness score between query and its relevant terms extracted from association rules (Luo et al. 2012). In this respect, after introducing some notations, we state the formal definitions of the concepts used in the remainder of the paper related to association rules and Explicit Semantic Analysis. Table 1 provides an overview of the notations used in this and the later sections.

Table 1 Summary of notations

Notation	Description
$\mathcal{C}$	The *whole set* of documents which form the collection
C	A *set* of documents belonging to the collection ($C \subseteq \mathcal{C}$)
d	A *single* document of the collection ($d \in \mathcal{C}$)
V	The *whole set* of *distinct* terms of the collection $\mathcal{C}$
T	A *set* of terms of the collection ($T \subseteq V$)
t	A *single* term of the collection ($t \in V$)
R	An association rule
q	An original query
t_q	A term in a given query q
E_q	A query q extended

In this paper, we represent a query q as set (bag) of terms, as follow:

$$q = \{t_{q1}, \ldots, t_{qn}\} \tag{1}$$

where t_{qi} is a term in a given query q and $i \in \mathbb{N}$.

3.1 Overview of extracting association rules from texts

Definition By analogy to the transactions used in data mining where each transaction is represented by ⟨Id-transaction, list-items⟩, we define a transaction in the text mining framework as follow: ⟨Id-document, list of terms contained in document⟩.

Basic formalism

Consider a set of n terms $V = \{t_1, t_2, \ldots, t_n\}$ and a corpus of m documents $C = \{d_1, d_2, \ldots, d_m\}$. Each d_i document included in C contains a subset of terms, T, included in V called termset.[3]

An association rule (R) binds two termsets, which respectively constitute its premise (T_1) and conclusion (T_2) parts (Agrawal and Skirant 1994). Thus, a R estimates the probability of having the terms of the conclusion (T_2) in a document, given that those of the premise (T_1) are already there. The advantage of the insight gained through association rules is in the contextual nature of the discovered inter-term correlations. Indeed, more than a simple assessment of pair-wise term occurrences, an association rule binds two sets of terms, which respectively constitute its premise and conclusion parts. Thus, a rule approximates the probability of having the terms of the conclusion in a document, given that those of the premise are already there.

Given a termset T, the support of T is equal to the number of documents in the document collection $\mathcal{C}$ containing all the terms of T. The *absolute* support of T is formally defined as follows (Han et al. 2000)[4]:

$$Supp(T) = |\{d | d \in \mathcal{C} \wedge \forall t \in T : (d, t) \in I\}| \tag{2}$$

The *relative* support of T is equal to:

[3] By analogy to the itemset terminology used in data mining for a set of items.

[4] In this paper, we denote by $|X|$ the cardinality of the set X.

$$\frac{Supp(T)}{|\mathcal{C}|} \tag{3}$$

where

– $I \subseteq \mathcal{C} \times T$ is a binary (incidence) relation. Each couple $(d, t) \in I$ indicates that the document d contains the term t.

A termset T is said to be *frequent* if its support value, i.e., supp(T), is greater than or equal to a user-defined threshold denoted *minsupp*.

A termset is said to be *closed* if none of its immediate supersets[5] has the same support as this original termset. Notice that in the remainder of the paper, we use the absolute support, i.e., Eq. (2).

Given a rule R: $T_1 \Rightarrow T_2$, the *support* of R is computed as follows:

$$Supp(R) = Supp(T_1 \cup T_2). \tag{4}$$

An association rule R is said to be *frequent* if its support value, i.e., supp(R), is greater than or equal to a user-defined threshold denoted *minsupp*. The *confidence* of R is computed as follows:

$$Conf(R) = \frac{Supp(T_1 \cup T_2)}{Supp(T_1)}. \tag{5}$$

An association rule R is said to be *valid* if its confidence value, i.e., $Conf(R)$, is greater than or equal to a user-defined threshold denoted *minconf*. This confidence threshold is used to exclude non valid rules.

3.1.1 Association rules extraction process

Given a set of n terms $V = \{t_1, t_2, \ldots, t_n\}$ and a collection of m documents $C = \{d_1, d_2, \ldots, d_n\}$ the extraction of association rules between terms, satisfying predefined thresholds of support *minsupp* and confidence *minconf*, is performed in two steps:

– A minimum support threshold is applied to find all frequent termsets in a documents collection. This phase consists in generating all termsets with a support greater than or equal to *minsupp*. These sets are called *frequent termsets*. The phase of generating frequent termsets is the most complex phase in the extraction process since it involves searching all possible termsets (term combinations), as in the context of exploring transactional databases (Agrawal et al. 1993). Several works in the literature are devoted to the study of so-called bibliometric laws or information law, which have been formulated from empirical observations on textual corpus. Among these laws, we cite the law of *Zipf* (Li 1992), that we have used in our work. *Zipf* law consists in studying textual corpus, regularities on the terms appearance frequency.
– A minimum confidence constraint is applied to these frequent termsets in order to form rules. Once the frequent termsets are derived, the generation of association rules is a fairly simple step. It is possible to adapt the algorithms for generating association rules from the frequent itemsets. In this work, we have adapted CHARM (Zaki and Hsiao 2002) algorithm to extract association rules.

[5] A, B two termsets, B is superset of A if $|A| < |B|$ and $A \subset B$.

3.1.2 CHARM algorithm

The CHARM algorithm was proposed by Zaki and Hsiao (2002), The originality of CHARM lies in the fact that it favors a depth-first exploration of the search space. Another characteristic to be mentioned to the credit of CHARM is that it uses a vertical representation, called diffset, to accelerate the calculation of the supports.

The task of mining association rules consists of two main steps, as we mentioned before. The first step involves finding the set of all frequent termsets. The second step involves testing and generating all high confidence rules among termsets. In CHARM, it is not necessary to mine all frequent termsets in the first step, instead it is sufficient to mine the set of closed frequent termsets which is much smaller than the set of all frequent termsets. It is also not necessary to mine the set of all possible rules.

In our case, considering as input the set of text documents, a minimal threshold of support *minsupp* and a minimal threshold of confidence *minconf*. This algorithm is able to derive all the association rules satisfying the constraint threshold of confidence of the rule.

An example of association rules is given in Table 2.

In the following, we define the association rules set mined from a collection $\mathcal{C}$ as follows:

$$\mathcal{R}_C = \{R | \, Conf(R) \geq minconf \;\; and \;\; Supp(R) \geq minsupp\} \tag{6}$$

3.2 Explicit semantic analysis (ESA)

Explicit Semantic Analysis (*ESA*) is a semantic relatedness measure proposed by Gabrilovich and Markovitch (2007). It is a technique that is somewhat reminiscent of the vector space model widely used in information retrieval. Here, documents are not represented by occurring terms but by their similarity to concepts derived from WIKIPEDIA articles. Each WIKIPEDIA concept is represented as an attribute vector of words that occur in the corresponding article. Entries of these vectors are assigned weights using $tf \times idf$ weighting. These weights quantify the strength of association between words and concepts. Thus, by comparing documents to all articles in a WIKIPEDIA corpus that has been preprocessed by tokenization, stemming, stop word removal and a term weight metric, a vector is obtained that contains a similarity value to each of the articles. A major advantage of ESA is that semantic relatedness can be calculated for terms and documents alike, providing good and stable results for both models.

Formally, the document collection is represented as an $n \times m$ matrix M, called *semantic interpreter*, where n is the number of articles and m the number of occurring terms in the corpus. M contains (normalized) $tf \times idf$ document vectors of the articles.

In order to evaluate the similarity between two given texts, terms or a text and a term, the cosine similarity measure is employed. In this paper, *ESA* is performed to compute the similarity between a query q and a term candidate t as follows:

$$ESA(q,t) = \frac{\overrightarrow{q} \times \overrightarrow{t}}{\left\|\overrightarrow{q}\right\| \times \left\|\overrightarrow{t}\right\|}. \tag{7}$$

where

– $\overrightarrow{q}$, $\overrightarrow{t}$ are the vectors generated by *ESA* that represent, respectively, the query q and the term t.

Table 2 Examples of association rules generated with CHARM (Zaki and Hsiao 2002) from WIKIPEDIA

R	Premise (T_1)	Conclusion (T_2)	$Supp(R)$	$Conf(R)$
Manufacture ⇒ car	Manufacture	Car	356	0.8921
Campus ⇒ university	Campus	University	279	0.7431
Manufacture motor ⇒ automobile car	Manufacture motor	Automobile car	143	0.7922

– $\|\overrightarrow{q}\|, \|\overrightarrow{t}\|$ are, respectively, the norm of the vectors $\overrightarrow{q}$ and $\overrightarrow{t}$.

In the next section, we introduce our HQE model which incorporates different external resources and combines local, global and external methods.

4 A hybrid query expansion model

In this section, we address the two basic issues of the QE process as detailed by Carpineto and Romano (2012), namely: terms generation and terms selection. In this respect, we propose to use multiple resources of knowledge, in addition to document collections, such as WIKIPEDIA and DBPEDIA to diversify expansion terms. More specifically, our model first automatically generates a list of candidate terms from each resource, and then combines the selected terms (using a semantic relatedness measure) for all the expanded queries. Our goal is to demonstrate the effectiveness of QE incorporating different resources coupled with a semantic selection.

In HQE, we propose three candidate term generation methods (*cf.* Fig. 1):

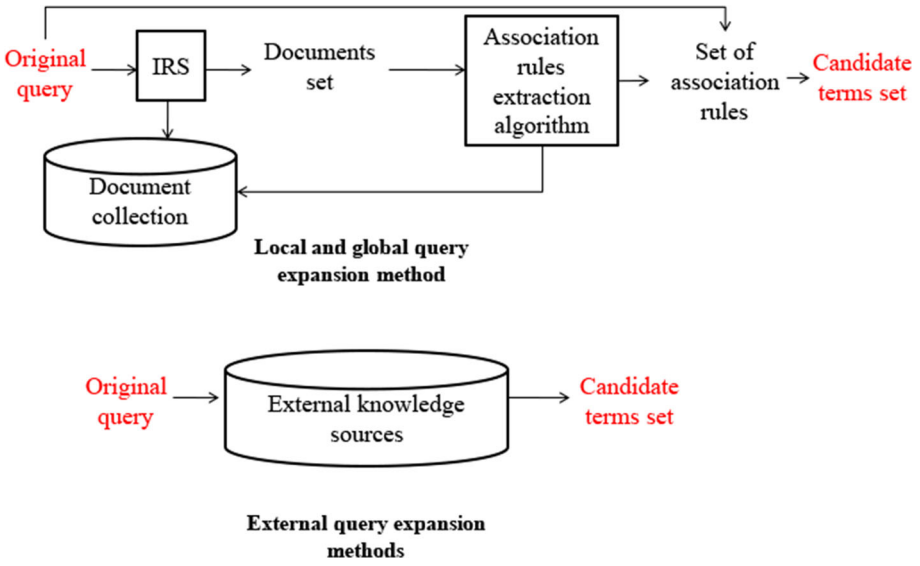

Fig. 1 Candidate terms generation

- The first one is based on a combination of local and global methods. Our hypothesis is that useful terms will occur in relevant documents more frequently than in non-relevant documents. Hence, we used a local method to retrieve, from a corpus, a set of relevant documents to the query. A global method is then used to generate the candidate terms by mining the association rules between terms from the retrieved documents;
- the two others use different external knowledge sources to generate the candidate terms.

As depicted in Fig. 2, the HQE process is split in two phases, namely:

- *Candidate terms generation* Phase 1 consists in generating the $Candidate\_Set(q)$ of a given query q;
- *Candidate terms selection* Phase 2 consists in (i) defining the $relatedness(q, t)$ function, (ii) ranking the $Candidate\_Set(q)$ set according to their relatedness score to the query returned by $relatedness(q, t)$; and (iii) selecting the best ones to be added.

These phases are detailed in the following.

4.1 Phase 1: candidate terms generation

Formally, given an original query $q = \{t_{q1}, \ldots, t_{qn}\}$, the set of candidate terms for its expansion is called $Candidate\_Set(q)$:

$$Candidate\_Set(q) = \{t_1, \ldots, t_p\} \tag{8}$$

where t_i is a candidate term.

We propose three different methods for this phase. The first one is a combination of local and global methods, and is called *statistical* query expansion method. Its objective is

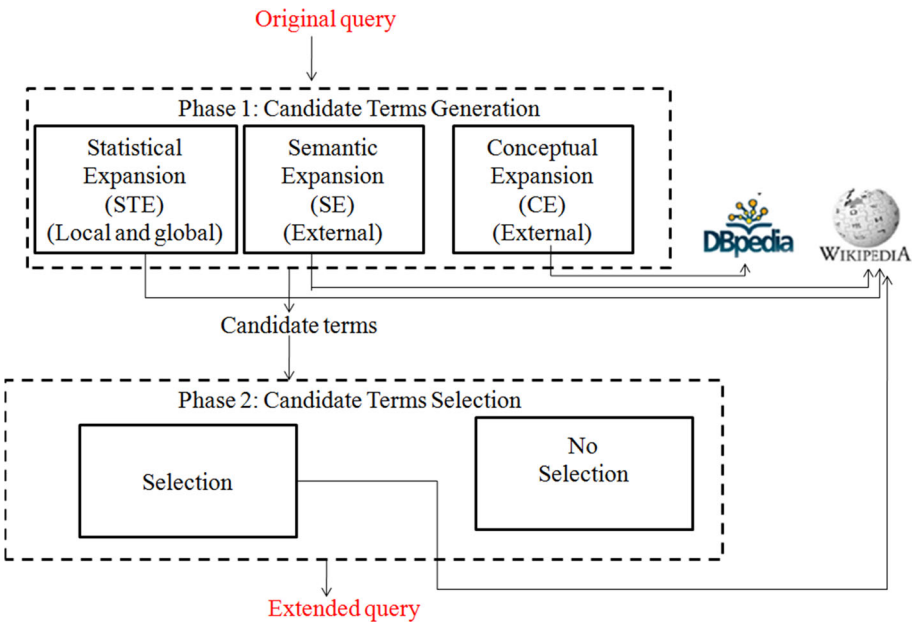

Fig. 2 The proposed hybrid model for query expansion

to generate, for a given query, statistically correlated terms without taking into consideration their semantics using association rules mining technique. The second method is an external method, and is called *semantic* query expansion method. It aims at generating new terms that are semantically related to the query. These terms are extracted from query's terms definitions. We note that this method takes into consideration the semantic aspect of the terms. The third one is also an external method, and is called *conceptual* query expansion method. Its goal is to generate new terms from an ontology by matching the query terms with the ontology concepts. We detail, in the following, these three methods of the proposed model.

4.1.1 Statistical expansion (STE)

The first method, denoted in the following STE, consists in retrieving a collection of documents C, in response to a given query, using an IRS. It consists in a local method followed by a global method: a PRF method since it uses the set of documents (C) retrieved from $\mathcal{C}$ in response to the original query, and association rules which are applied to select candidate terms.

The collection of documents (C) is used to extract a set of association rules ($\mathcal{R}_C$) that discover strong correlations between terms. The set of candidate terms generated by STE ($Candidate\_Set_{STE}(q)$) is built upon all the association rules of $\mathcal{R}_C$ having their premise included in q, as[6]:

$$Candidate\_Set_{STE}(q) = \bigcup_{(T_1 \Rightarrow T_2) \in \mathcal{R}_C \text{ sothat } T_1 \in 2^q} T_2 \tag{9}$$

The process of generating the candidate terms set $Candidate\_Set_{STE}(q)$, for a given query q, is performed in the following steps:

- Selecting a collection of documents C, similar to the query, using an IRS. We used texts extracted from WIKIPEDIA articles as a documents collection. We get, using the Terrier system, the top-k in answers from the set of queries to generate C;
- Collection annotating: in order to extract the most representative terms, a linguistic preprocessing is performed on C by using a part-of-speech tagger TREETAGGER.[7] In this work, we keep the common nouns and the proper nouns, since they are the most informative grammatical categories and are most likely to represent the content of documents (Barker and Cornacchia 2000). A stoplist is used to discard functional English terms that are very common;
- Mining the association rules from C expressing the correlations between terms using the algorithm CHARM of Zaki and Hsiao (2002);
- Generating the candidate terms set, $Candidate\_Set_{STE}(q)$, using the association rules mined, for q.

4.1.2 Semantic expansion (SE)

This method, denoted in the following SE, consists in extending the queries using semantically related terms that come from an external structured semantic source of

[6] With 2^X denoting the set of all subset of the set X.

[7] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

knowledge called *RS*. We assume that the documents in this resource are structured: they are textual documents that are dedicated to describe concepts represented by terms, and their structure contains a part dedicated to the definition of the concept. We define the $Def_{Semantic}(t, RS)$ function that returns, for a given term t, *a semantic definition* extracted from the documents of *RS* as follows:

$$Def_{Semantic}(t, RS) = \{t_1, \ldots, t_k\} \tag{10}$$

The set of candidate terms generated by SE ($Candidate\_Set_{SE}(q)$) is defined as follows:

$$Candidate\_Set_{SE}(q) = \bigcup_{t \in q} Def_{Semantic}(t, RS) \tag{11}$$

To achieve this, we use some heuristics:

- First, given a query, we search all documents from *RS* that correspond to the query's terms.
- We extract, from these documents, the corresponding definitions;
- We annotate these definitions by applying the same annotating process as described in Sect. 4.1.1, then, we extract a set of specific terms (i.e., nouns), that are the candidate terms to expand the original query.

For this QE method, we opted for WIKIPEDIA as the semantic knowledge source *RS*. This latter has the following significant features: wide knowledge coverage, rich semantic knowledge, highly structured and high speed of information update. Therefore, it is an ideal data resource to improve a QE process (Gan and Hong 2015). In our case, since the Wikipedia articles follow a predictable layout, the definition of an article[8] is the article's first sentence and paragraph.

4.1.3 Conceptual expansion (CE)

This method, denoted CE, relies on an external ontology O to extract related concepts to a given term t in the original query. We define the function $Concept(t, O)$ that returns the related concepts to t from the ontology O as follows:

$$Concept(t, O) = \{t_1, \ldots, t_P\} \tag{12}$$

The set of candidate terms generated by CE ($Candidate\_Set_{CE}(q)$) is defined as follows:

$$Candidate\_Set_{CE}(q) = \bigcup_{t \in q} Concept(t, O) \tag{13}$$

For the CE method, we used DBPEDIA as an ontology O. It is an ontology extracted from WIKIPEDIA and aims to represent the WIKIPEDIA content in Resource Description Framework (RDF) triples.

The process of generating the candidate terms set $Candidate\_Set_{CE}(q)$, for a given query q, is performed in the following steps:

[8] Articles supply the bulk of Wikipedia's informative content. Each article describes a single concept or topic.

- First, we project the query terms on the ontology concepts. This matching is done using SPARQL[9];
- We leverage the descriptions (*rdf:type*) of the concepts as each description of a concept may be related words, synonyms, or alternative terms that refer to the concept;
- We use these descriptions to extend the original query.

4.2 Phase 2: candidate term selection

Roughly speaking, the relatedness function returns the relatedness score between a query q and a candidate term $t \in Candidate\_Set(q)$:

$$relatedness(q,t) = score \in \mathbb{R} \tag{14}$$

The term t is considered relevant with respect to the query q if and only if $score \geq \mu$, where μ is the minimal threshold of the semantic relatedness.

Consequently, the extended query is provided by selecting the most related terms among the $Candidate\_Set(q)$ as sketched in Eq. (15):

$$E_q = q \cup \{t \in Candidate\_Set(q) \mid relatedness(q,t) = score \geq \mu\} \tag{15}$$

To define the *relatedness* function, we propose a new measure (denoted *E*SAC) that combines, using a linear interpolation method, the WIKIPEDIA-based Explicit Semantic Analysis (*ESA*) (*cf.* Sect. 3.2) measure and the association rules' confidence one's. In the case when a term has a non-0 ESA value with the query but no association rule is available between this term and the query, we choose to keep the ESA score only, in a way to avoid over-penalizing it. *ESAC* is defined as follows:

$$\begin{aligned} relatedness(q,t) &= ESAC(q,t) \\ &= \begin{cases} (\alpha \times ESA(q,t) + (1-\alpha) \times Conf_{max}(R,q,t) & if\ Conf_{max}(R,q,t) \neq 0; \\ ESA(q,t), & otherwise. \end{cases} \end{aligned} \tag{16}$$

where

$$Conf_{max}(R,q,t) = \max_{t_q \in q, R \in \mathcal{R}_C} Conf(R(t_q,t)) \tag{17}$$

is the maximum of the confidence of any association rule R in $\mathcal{R}_C$, any term t_q from the query q, and the candidate term t.

4.3 Configurations for our HQE model

Since we propose three different methods for the *Candidate terms generation phase*, namely, STE, SE and CE that generate for each query q different sets of candidate terms, and two alternatives for the selection phase based on the ESAC measure, we can derive different configurations of the expansion term sets as depicted in Table 3.

From the configurations identified in Table 3, we get:

[9] http://dbpedia.org/sparql.

Table 3 The different configuration of our HQE model

Terms generation	Terms selection	
	With selection	Without selection
STE	$\text{STE}_{Selection}$	$\text{STE}_{NoSelection}$
SE	$\text{SE}_{Selection}$	$\text{SE}_{NoSelection}$
CE	$\text{CE}_{Selection}$	$\text{CE}_{NoSelection}$
ALL	$\text{ALL}_{Selection} = \text{STE}_{Selection} \cup \text{SE}_{Selection} \cup \text{CE}_{Selection}$	$\text{ALL}_{NoSelection} = \text{STE}_{NoSelection} \cup \text{SE}_{NoSelection} \cup \text{CE}_{NoSelection}$

Capital letters denote the corresponding runs in the experimental validation

- $\mathbf{SE}_{NoSelection}$ the set of the expansion terms stems from applying the semantic expansion (SE) method, aforementioned, without considering the selection phase: $E_q = q \cup Candidate\_Set_{SE}(q)$;
- $\mathbf{SE}_{Selection}$ the set of the expansion terms is derived by applying the semantic expansion (SE) method, aforementioned, involving the selection phase: $E_q = q \cup \{t \in Candidate\_SetSE(q) \mid relatedness(q,t) = score \geq \mu\}$;
- $\mathbf{STE}_{NoSelection}$ the set of the expansion terms is generated by applying the statistical expansion (STE) method, aforementioned, without any selection phase: $E_q = q \cup Candidate\_Set_{STE}(q)$;
- $\mathbf{STE}_{Selection}$ the set of the expansion terms are generated by applying the statistical expansion (STE) method, aforementioned, and the selection phase: $E_q = q \cup \{t \in Candidate\_Set_{STE}(q) \mid relatedness(q,t) = score \geq \mu\}$;
- $\mathbf{CE}_{NoSelection}$ the set of the expansion terms stems from applying the conceptual expansion (CE) method, aforementioned, without any selection phase: $E_q = q \cup Candidate\_Set_{CE}(q)$;
- $\mathbf{CE}_{Selection}$ the set of the expansion terms stems from applying the conceptual expansion (CE) method, aforementioned, and the proposed selection phase: $E_q = q \cup \{t \in Candidate\_Set_{CE}(q) \mid relatedness(q,t) = score \geq \mu\}$;
- $\mathbf{ALL}_{Selection}$ the set of the expansion terms is the conjunction of the following sets: $\mathrm{SE}_{Selection}$, $\mathrm{STE}_{Selection}$ and $\mathrm{CE}_{Selection}$: $E_q = q \cup \{t \in Candidate\_Set_{SE}(q) \cup Candidate\_Set_{STE}(q) \cup Candidate\_Set_{CE}(q) \mid relatedness(q,t) = score \geq \mu\}$;
- $\mathbf{ALL}_{NoSelection}$ same as the $\mathrm{ALL}_{Selection}$ without any selection phase.

The remainder of this paper is devoted to the experimental validation of the HQE model in two different tasks namely: ad-hoc IR and Tweet contextualization.

5 Experimental validation in IR

In order to prove the effectiveness of our HQE model, we focus on two cases in which the potential mismatch between the queries expression and the documents has to be tackled, namely: (i) a social IR task dealing with tweet search using TREC Microblog 2011 test collection, in which both the queries and the tweets are short; and (ii) an ad-hoc IR task using TREC Robust 2004 collection in which the queries are known to be difficult due mainly to term mismatch. We study the performances of the different configurations of our HQE model and compare them with classical PRF approach.

5.1 Test collections description

The following are the details of the two considered collections:

- The **TREC 2011 Microblog Track** is a *social text collection* that addresses a real-time search task, where the user looks for the most recent but relevant information (tweets) to the query. The TREC Microblog 2011 test collection contains 16 million tweets collected over a period of 2 weeks (Ounis et al. 2011). This task proposes 50 topics with no narrative and description tags are provided, so this collection is a good test for our proposal. Contrary to the task, our concern here is not to rank results according to time but to relevance, this is why we do compare our configurations, but we do not put our results in perspective to the official ones of TREC 2011 Microblog Track.

Table 4 Cross validation parameter values per model

Model	Parameter	Values
BM25	b	0.01; 0.02; 0.03; 0.04; 0.05; 0.06; 0.07; 0.08; 0.09; 0.1; 0.2, 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9
Hiemstra	λ	0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9
Dirichlet	μ_D	50; 300; 500; 1000; 1500; 2000; 2500; 3000

However, to keep our run consistent with the assessments of this evaluation, we remove from our results the tweets with timestamps later than the query.

– The **TREC 2004 Robust retrieval track**, Voorhees (2004), evaluates ad-hoc retrieval. This test collection includes four document collections composed of administrative documents, and English newspaper articles, with a total of more than a half of a million documents. As stated before, our concern is to study our proposal in the case of difficult queries, this is why we chose to focus only on a subset of the queries called *hard topics*. These 50 topics, known to be difficult for current automatic systems, were drawn by the track organizers from topics 301–450 that were used in the TREC 2003 robust track. We choose to focus on this task because Voorhees (2004) states that all of the top-performing runs [like Kwok et al. (2004)] in TREC 2004 Robust retrieval track used the web to expand queries. Because we also propose a framework that relies on external Web data, such experiment makes sense.

5.2 Experimental setup

All the experiments reported here are achieved on the Terrier 4.0 IR platform (Ounis et al. 2005). We test our query extension configurations with classical IR models, namely: BM25, language models with Dirichlet or Hiemstra (i.e., Jelinek-Mercer) smoothings, and we also consider the classical PRF approach. The IR models parameters are optimized using a fivefold cross-validation on sets of queries separated from the tasks queries (topics) described above: for the TREC microblog task, optimized on the INEX collection (Bellot et al. 2016) (see Sect. 6) as we assume similar behavior on both of these corpus; for the hard topics of the Robust track, the optimization is achieved on the 249 non-hard topics of TREC 2004 Robust Track. Table 4 resumes the IR models, the set of parameters and their range values used for optimization.

In order to build association rules between terms for our TREC microblog task 2011 experiments, the tweets are not a good source, so we used, overall, a set of 50,000 WIKIPEDIA articles. For the experiments on the TREC Robust Task 2004, we build the rules using the whole corpus. The rules are defined using the CHARM algorithm by Zaki and Hsiao (2002). The minimal threshold of the support, *minsupp*, is experimentally set as follows: we varied the minimum and maximum threshold of the support, i.e., *minsupp* and *maxsupp*,[10] w.r.t. the document collection size and term distributions. While considering the *Zipf* distribution of every collection, the maximum threshold of the support values is experimentally set in order to spread trivial terms which occur in most documents, and are then related to too many terms. On the other hand, the minimal threshold allows eliminating marginal terms which occur in few documents, and are then not statistically important when occurring in a

[10] *maxsupp* means that the termset must occur at most below than this user-defined threshold.

Table 5 Description of the IR test collections considered

Collection	#documents	*Minsupp*	*Minconf*	#Rules
WIKIPEDIA corpus for TREC Microblog 2011 (50 social topics)				
Documents	50,000	15	0.7	402,862
TREC 2004 Robust Track (50 hard topics)				
FBIS	130,471	1,700	0.5	770,359
Federal register 94	55,630	1,500	0.5	211,759
LA times	131,896	2,000	0.5	538,323
Financial times	210,158	2,500	0.5	379,248

Table 6 ESAC parameters ranges for optimization

Parameter	Values
α	0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9
μ	0.35; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9

rule. In the same way, the minimal threshold of the confidence of mined association rules *minconf* is defined by varying it between 0.4 and 1.0 by a step of 0.1. The best values of these thresholds, i.e., absolute *minsupp* and *minconf* respectively for the different collections are given in Table 5.

For the selection of the most related terms to the original query, the parameters α and μ of the semantic relatedness function ESAC [*cf.* Eqs. (15, 16)] are optimized using fivefold cross-validation according to the MAP evaluation measure using the BM25 IR model on the same set of queries as the other optimizations. Table 6 lists the set of parameters, and the values used in the experiments.

Based on the different experiments conducted, the optimal values for α and μ are stable, this is why we set in all of our experiments these parameters to $\alpha = 0.5$ and $\mu = 0.4$.

5.3 Experiments and results for tweets search (TREC 2011)

TREC 2011 Microblog Track results show that Tweets retrieval is far from being a solved problem, we expect that our proposals will improve the quality results for this task. Once again, we underline the fact that our goal does not fit into the official evaluation of results related to timeline, we do not then compare our results to the official ones of these evaluation campaign. The obtained results are depicted in Table 7, where, the results are ranked by performance on MAP for each of the three IR models studied. We present the precision at 5, 10 and 30 documents, the Mean Average Precision (MAP), as well as the percentage of the MAP increase according to the respective baseline of each IR model.

From Table 7 we see that:

– For each model, all our query expansion configurations (filtered or not) outperform the non-expanded Baselines, as well as the classical Pseudo Relevance Feedback. PRF on very short documents does not seem to be a good solution. Thus, our framework that relies on external information is clearly beneficial.

Table 7 Comparative evaluation of retrieval effectiveness for TREC Microblog 2011

Run	Configuration	P@5	P@10	P@30	MAP (%Chg.$_{Baseine}$,%Chg.$_{PRF}$)
BM25					
Baseline	–	0.1265	0.1327	0.1238	0.1025
PRF	–	0.1592	0.1551	0.1245	0.1145
–	**STE$_{Selection}$**	**0.4000**	**0.3796**	**0.3197**	**0.3079** $^{\dagger\circ}$ (**200%** , **168%**)
–	STE$_{NoSelection}$	0.3551	0.3265	0.2850	0.2804$^{\dagger\circ}$ (173% , 145%)
–	ALL$_{Selection}$	0.3633	0.3429	0.2707	0.2747$^{\dagger\circ}$ (168% , 140%)
–	SE$_{Selection}$	0.3342	0.3184	0.2626	0.2589$^{\dagger\circ}$ (153% , 126%)
–	ALL$_{NoSelection}$	0.3551	0.3388	0.2553	0.2570$^{\dagger\circ}$(151% , 124%)
–	CE$_{Selection}$	0.3224	0.3041	0.2755	0.2505$^{\dagger\circ}$(144% , 118%)
–	CE$_{NoSelection}$	0.2408	0.2227	0.2041	0.2053$^{\dagger\circ}$ (100% , 79%)
–	SE$_{NoSelection}$	0.2367	0.2224	0.2163	0.1676$^{\dagger\circ}$(63% , 46%)
HIEMSTRA					
Baseline	–	0.1429	0.1429	0.1333	0.1148
PRF	–	0.1469	0.1755	0.1374	0.1156
–	**STE$_{Selection}$**	**0.3837**	**0.3673**	**0.3014**	**0.3083** $^{\dagger\circ}$ (**168%**, **166%**)
–	STE$_{NoSelection}$	0.3469	0.3347	0.2939	0.2883$^{\dagger\circ}$(151%, 149%)
–	CE$_{Selection}$	0.3306	0.3286	0.2653	0.2690$^{\dagger\circ}$(134%, 132%)
–	SE$_{Selection}$	0.3184	0.3102	0.2605	0.2627$^{\dagger\circ}$(128%, 127%)
–	CE$_{NoSelection}$	0.2857	0.2755	0.2265	0.2439$^{\dagger\circ}$(112% , 110%)
–	ALL$_{Selection}$	0.3102	0.2796	0.2265	0.2177$^{\dagger\circ}$ (89% , 88%)
–	ALL$_{NoSelection}$	0.3020	0.2714	0.2286	0.2075$^{\dagger\circ}$(81%, 79%)
–	SE$_{NoSelection}$	0.2245	0.2286	0.1986	0.1671$^{\dagger\circ}$ (45% , 44%)
DIRICHLET					
Baseline	–	0.1592	0.1571	0.1367	0.1156
PRF	–	0.1184	0.1286	0.1340	0.1177
–	**STE$_{Selection}$**	**0.4000**	**0.3837**	**0.3197**	**0.3152** $^{\dagger\circ}$ (**172%** , **167%**)
–	STE$_{NoSelection}$	0.3592	0.3367	0.3061	0.2973$^{\dagger\circ}$(157% , 152%)
–	CE$_{Selection}$	0.3540	0.3286	0.2762	0.2741$^{\dagger\circ}$ (137%, 132%)
–	SE$_{Selection}$	0.3184	0.3122	0.2741	0.2700$^{\dagger\circ}$(133% , 129%)
–	CE$_{NoSelection}$	0.2980	0.2857	0.2510	0.2507$^{\dagger\circ}$ (116%, 112%)
–	ALL$_{Selection}$	0.3347	0.2796	0.2354	0.2377$^{\dagger\circ}$(105%, 101%)
–	ALL$_{NoSelection}$	0.3020	0.2714	0.2286	0.2075$^{\dagger\circ}$(79%, 76%)
–	SE$_{NoSelection}$	0.2163	0.2122	0.2000	0.1739$^{\dagger\circ}$ (50% , 47%)

Bold values indicate the best results

$\%Chg._{baseline}$, $\%Chg._{PRF}$ indicate the model improvements in terms of MAP compared, respectively, to the baseline and PRF. The symbols † and ∘ denote significant MAP difference based on, respectively, the baseline run and PRF run (*t*-test, $p \leq 0.05$)

– Consistently, the configurations with terms selection outperform their un-filtered counterparts. This shows that our filtering proposal based on the *ESAC* measure is beneficial for such a task;

Table 8 Percentage of queries R^+, R^- and $R^=$ for which $STE_{Selection}$ performs better, lower, equal to, the BM25 baseline and PRF BM25, for the MAP

Query set	#Queries (%)
BM25 baseline	
R^+	46/49 (93, 87%)
$R^=$	0/49 (00.00%)
R^-	3/49 ((06.12%)
PRF BM25	
R^+	42/49 (85.71%)
$R^=$	0/49 (00.00%)
R^-	7/49 (14.28%)

- For the three models, the best configuration is the one that relies on the statistical expansion using association rules, $STE_{Selection}$. The integration of all extensions also gives good results (third best) for the BM25 model;
- The fifth best result for BM25 uses all extensions without any filtering: it seems then that BM25 is better at managing the noisier queries.

Bilateral paired *Student t-test* (Smucker et al. 2007) on MAP results, comparing on one hand the baseline with our runs, and these latter with the PRF run on the other; where the results show that the differences are almost always significant with $p < 0.05$ (*cf.* Table7).

For further analysis of the effectiveness, we present in Table 8 a gain and failure analysis of the HQE model. It presents statistics on the queries, for R^+, R^-, and $R^=$ for which $STE_{Selection}$ performs better (subset R^+), lower (subset R^-), and equal (subset $R^=$) than the baseline and PRF, in terms of MAP. From Table 8, we see that our best configuration is clearly better, more than 85% of queries for the PRF and almost 94% of queries for the un-extended ones.

We conclude from this experiments on the TREC microblog that, i.e., from Tables 7 and 8, our query expansion leads to the improvement of the retrieval effectiveness in the case of short queries and short documents.

5.4 Experiments and results for TREC 2004 robust retrieval track

We conducted experiments using one or more elements of the topics, namely: the title, description and narrative fields. Using all of these fields is expected to boost the quality of the results. When conducting the experiments, we found that: (i) the language models were outperformed by BM25 (due to the fact that, as noted above, BM25 seems more keen to lower the impact of un-related terms), that is why Table 9 presents only BM25 results, and (ii) that our best results were obtained with **Title+Description+Narrative** runs, so only these runs are reported here. Again, in this table, the results of our configurations are ranked according to the MAP values. At the bottom of Table 9, we present the best official result [run id pircRB04td2,[11] by the City University of New York (Kwok et al. 2004)] from TREC 2004 Robust track hard topic set: we selected the run with the higher MAP in among the official runs. This run considers title+description topics. Unofficial results[12] at TREC 2004 Robust track from the National Laboratory of Pattern Recgnition (Beijin, China) used title+description+narrative, but they were all outperformed by the pircRB04td2 run.

[11] http://trec.nist.gov/pubs/trec13/appendices/robust/pircRB04td2.table.pdf.

[12] http://trec.nist.gov/pubs/trec13/appendices/robust/NLPR04okall.table.pdf.

Table 9 Comparative evaluation of retrieval effectiveness for Robust TREC 2004 (Hard topics) under BM25 Model

Run	Configuration	P@5	P@10	P@30	MAP($\%Chg_{\cdot baseline}$, $\%Chg_{\cdot PRF}$)
Title+Description+Narrative					
Baseline	–	0.3760	0.3200	0.2033	0.1339
PRF	–	0.4240	0.3520	0.2633	0.1546
–	$STE_{NoSelection}$	0.4160	0.3640	0.2780	$0.1471^{\dagger}$ (+ 10% , − 4%)
–	$ALL_{NoSelection}$	0.3640	0.3600	0.2680	$0.1453^{\dagger}$ (+ 9% , − 6%)
–	$ALL_{Selection}$	0.3760	0.3480	0.2687	$0.1422^{\dagger}$ (+ 6% , − 8%)
–	$CE_{Selection}$	0.3840	0.3460	0.2587	$0.1418^{\dagger}$ (+ 6% , − 8%)
–	$STE_{Selection}$	0.3800	0.3440	0.2613	$0.1403^{\dagger}$ (+ 4%, − 9%)
–	$SE_{Selection}$	0.3360	0.3060	0.2433	0.1395 (+ 4% , − 9%)
–	$SE_{NoSelection}$	0.3320	0.3060	0.2373	0.1353 (+ 1% , − 12%)
–	$CE_{NoSelection}$	0.3640	0.3240	0.2560	0.1352 (0%, − 12%)
PRF	$STE_{NoSelection}$	0.4120	0.3620	0.2867	$0.1777^{\dagger\circ}$ (+ 33% , + 14%)
Official best (pircRB04td2)		0.4600	0.4020	0.2867	0.1949

$\%Chg_{\cdot baseline}$, $\%Chg_{\cdot PRF}$ indicate the model improvements in terms of MAP compared, respectively, to the baseline and PRF. The symbols † and ∘ denote significant MAP difference based on, respectively, the baseline run and PRF run (*t*-test, $p \leq 0.05$)

Considering more classical Title only runs, PRF is outperformed by our best Title only run, $STE_{NoSelection}$, for P@5 P@10, whereas it outperforms our best proposal $STE_{NoSelection}$ for P@30 and MAP evaluation measures. This shows that our proposal is more precise, as it clearly generates better expansion terms for the top 5 and top 10 results. We chose in the paper to present our best results according to the available topic data, so we focus on Title+Description+Narrative runs.

We see in Table 9 that:

- The top-2 configurations (according to the MAP) for the hard topics are $STE_{NoSelection}$ and $ALL_{NoSelection}$, aka the configurations that use large expansions: the statistical expansion, or the union of all the expansion terms, without filtering. This shows again that our statistical expansion proposal, used alone or with other expansions, is effective. Another element related to the interest of STE configurations is that $STE_{NoSelection}$ obtains the third best result in P@30, with a value of 0.2780. In the case of difficult queries, it seems that integrating association rules is positive when retrieving the top documents;
- Most of the time, except for STE and ALL expansions, the filtering of expansion terms leads to better results. Our explanation for this is that: (i) the terms obtained by STE are already adequate, and the filtering does remove useful terms and, (ii) because of the unions of expansions with "ALL", the decrease of quality of STE also degrades the overall union of expansion terms;
- The PRF is very effective for MAP because it is applied on large documents. However, for the precision at 30 documents our three best configurations obtain better results;
- Even if our results according to the MAP do not attain the best official results, we see that our best precision value at 30 documents, 0.2780, is obtained by our statistical expansion, which is close to the official best of TREC 2004, (0.2867).

- Although the PRF method gives some very good results in hard queries, some of our runs have close results with those of PRF. Furthermore, a coupling of our run ('$STE_{NoSelection}$') and PRF gave much better results from our previous runs and those of PRF alone. Thus, we have managed to find a more efficient method in using '$STE_{NoSelection}$ + PRF', for these hard queries.
- All of our results are outperformed by the best TREC 2004 robust track run, pircRB04td2, on hard queries. We however have to mention that this run relies on fusing multiple (up to 4) retrieval lists from several queries, where our framework is limited to one expended query.

An additional experiment (reported in the penultimate line of Table 9), applied PRF on the extended **STE** $_{NoSelection}$ expansion. Such integration of complementary statistical query expansion and pseudo relevance feedback improves the recall after the top-30 documents, leading to a higher MAP value.

The results presented in Table 9 show that our approach leads to good results in terms of precision at 30 documents concerning hard topics. For this test collection, the filtering of terms plays a positive role in the semantic and conceptual expansions, as it filters out inadequate terms, but for STE the impact of the filtering is negative. This may be explained by the fact that many documents of the TREC Robust track are journal articles, and in with such *clean* documents the associations rules are already very precise, without the need for subsequent filtering. The good results obtained at 30 documents make us believe that when such effectiveness is needed, our proposals improve results, and such idea is confirmed in the following experiments related to tweet contextualization.

6 Embedding hybrid query expansion model for tweet contextualization

Microblogging platforms such as Twitter are increasingly used for online client and market analysis. This motivated the proposal of a new track at CLEF INEX lab of Tweet Contextualization in 2013. The objective of this task was to help a user to understand a tweet by providing a short explanatory summary. This summary should be built automatically using resources like WIKIPEDIA and generated by extracting relevant passages and aggregating them into a coherent summary (Bellot et al. 2016).

The general process of the tweet contextualization task classically involves three steps:

- Tweet analysis in a way to define what the tweet is about;
- Document retrieval, in a way to gather additional information that will serve as a basis for the contextualization;
- Summary generation, to generate an overview that describes the tweet. For INEX, the summary is not larger than 500 words.

INEX organizers provide the task participants with a baseline system that is composed of an IRS (based on the INDRI[13] search engine), and an automatic summarization system which takes as input a text composed of the top 50 results obtained by the IRS system

[13] http://www.lemurproject.org/indri.php.

Table 10 Description of INEX 2014 and 2013 test collections

Collection	$Source_{doc}$	#docs	$Source_{tw}$	#topics (tweets)
INEX 2014	English WIKIPEDIA from November 2012	3,902,346	Twitter (from CLEF RepLab 2013)	598
INEX 2013	"	"	Twitter	240

(summarization algorithm created by TERMWATCH[14] Ibekwe-Sanjuan and SanJuan 2004). This allows participants to focus on the best tweet formulations for the information retrieval system.

Since the quality of tweets have a direct impact on the contexts quality, we propose to use our proposed HQE model to enhance the tweets quality for the baseline system. As described above, we know that our proposed query extension configurations lead to high quality results for tweet retrieval. We also show that in the case of difficult queries our proposals obtain good results, especially when considering top 30 or more documents. Our goal here is to study the effectiveness of our query expansion model, and to evaluate the results using the INEX baseline system, on the INEX 2014 and 2013 test collections. We focus on the INEX 2014 campaign first, because we participated officially in it and we obtained the top result with our $STE_{NoSelection}$ configuration. Then, we describe post INEX 2014 experiments, as well as results on the INEX 2013 topics. In the following experiments reported, we focus only on the textual part of the tweets, without integrating specific elements like *hashtags*, *urls* or *mentions*.

6.1 INEX test collections

6.1.1 Data

The INEX 2014 and INEX 2013 collections are described in the Table 10. The documents corpus is the same (3 millions WIKIPEDIA english articles, notes and bibliographic references removed), and the topics (tweets) to be contextualized are different.

These two collections differ by their topics only:

- The INEX 2014 topic tweets are extracted from CLEF Replab 2013 which is dedicated to reputation monitoring. Each of these 240 tweets mention explicitly a company (e.g., *Fiat, Goldman Sachs,etc.* or an institution (e.g., Bank of America, New York University, etc.). So, using several ways to access sources of information that depicts such entities may lead to good results. Moreover, as many topic tweets are related to the same entities but differ according to the details discussed in the tweets, filtering the query expansion terms is also expected to provide good results;
- Compared to the 2014 topics, the INEX 2013 topic tweets ones are much obviously related to Wikipedia pages (for instance "Bulgaria's prime minister says he and his whole government is resigning from office following nationwide protests—@Reuters"). Moreover, many tweets are strongly time-related (the tweet above related to an event that took place in February 2013): it is difficult to get much relevant information when considering a WIKIPEDIA dump of 2012. Other tweets are quite obscure, like "But

[14] http://data.termwatch.es.

from each crime are born bullets that will one day seek out in you where the heart lies.—Pablo Neruda": what context is relevant to a tweet, even for human being, is not clear. We will see in Table 16 that the best results are much lower compared to INEX 2014. That is why we expect our proposal to behave also less well.

6.1.2 Evaluation metric

The official evaluation metric is the *Informativeness*, Bellot et al. (2016), which is not a classical evaluation metric for IR. Its goal is to measure how well the summary helps a user understand the tweets content (Bellot et al. 2016).

The informativeness of a given summary is the dissimilarity between this latter and a reference summary.

There are different distributions for the reference summaries, namely:

- Unigrams made of single lemmas (after removing stop-words).
- Bigrams made of pairs of consecutive lemmas (in the same sentence).
- Bigrams with 2-gaps also made of pairs of consecutive lemmas but allowing the insertion between them of a maximum of two lemmas (also referred to as skip distribution).

These distributions are more or less strict: Unigrams only consider simple overlap, where Bigrams take into account successive words.

It is important to remember for the analysis of the results that this evaluation metric is good when it is low.

6.2 Results on INEX 2014 test collection

6.2.1 Official INEX 2014 results

Before describing the official results we obtained at INEX 2014, we present in the second column of Table 11 the size of the WIKIPEDIA corpus used to build the association rules, and the other columns give the *minsupp* and *minconf* parameters used, as well as the overall number of rules generated for the STE configurations.

We note that we used the same values for α and μ as in the previous experiments.

Table 12 presents our official run submitted to INEX 2014 tweet contextualization track. It is worth noting that our run achieved the best result (Bellot et al. 2014), This run corresponds to the configuration $\text{STE}_{NoSelection}$ described in our previous experiments, using the statistical expansion by association rules with parameters presented in Table 11. This shows again that our statistical expansion is successful: the expansion is then able to retrieve WIKIPEDIA pages that lead to higher quality contexts. Based on such results, we describe in the following section, additional experiments to study the other configurations proposed in Sect. 4.3.

Table 11 Description of WIKIPEDIA corpus extracted for 2014 Tweet Contextualization

Collection	#docs	*Minsupp*	*Minconf*	#Rules
Documents	50,000	15	0.7	378,212

Table 12 INEX Tweet Contextualization 2014 official informativeness results (Bellot et al. 2016)

Run id.	Unigrams	Bigrams	Bigrams with 2-gaps
361 ($STE_{NoSelection}$)	0.7632	0.8689	0.8702

Table 13 The obtained informativeness results on INEX 2014 collection

Configuration	Unigrams	Bigrams	Bigrams with 2-gaps
$ALL_{Selection}$	0.7494	0.8520	0.8535
$SE_{Selection}$	0.7613	0.8629	0.8630
$CE_{Selection}$	0.7610	0.8629	0.8638
$SE_{NoSelection}$	0.7665	0.8661	0.8668
$STE_{Selection}$	0.7612	0.8671	0.8695
$STE_{NoSelection}$ (361)	0.7632	0.8689	0.8702
$CE_{NoSelection}$	0.7940	0.8822	0.8831

6.3 Post INEX 2014 experiments

6.3.1 Configurations results

All our proposed configurations applied on the test collection of INEX 2014 lead to the results presented in Table 13.[15] Table 13 highlights our last results, ranked by *Informativeness* on Bigrams with 2-gaps.

The analysis of Table 4.3 shows that:

- The configuration that integrates the three resources, namely $ALL_{Selection}$, has achieved the best informativeness results for the three evaluation measures. This shows that the terms coming from the three terms generations are complementary;
- Consistently with the experiments on TREC microblog 2011, the filtered configurations $SE_{Selection}$, $STE_{Selection}$ and $CE_{Selection}$ outperform their respective un-filtered counterparts $SE_{NoSelection}$, $STE_{NoSelection}$ and $CE_{NoSelection}$. This shows that the filtering we propose is also valuable for tweets contextualization;

Overall, the framework proposed, which integrates several sets of terms from several sources, combined with an adequate filtering, enhance the quality of the generated context of the tweets.

6.3.2 Significance and effectiveness evaluation

To provide an in-depth understanding of the configuration $ALL_{Selection}$ improvement in comparison with $STE_{NoSelection}$, we present in Table 14 the percentage of queries R^+ and R for which the $ALL_{Selection}$ configuration performs better (worse) than $STE_{NoSelection}$, with Bigrams with 2-gaps evaluation informativeness measure, in terms of informativeness metric . For 27 topics, $ALL_{Selection}$ outperforms $STE_{NoSelection}$, and for 27 topics it underperforms $STE_{NoSelection}$. It shows that it is difficult to choose between these configurations.

[15] For post INEX 2014 experiments, the runs $ALL_{NoSelection}$ could not be tested because the system provided by the INEX organizers is no more accessible.

Table 14 Percentage of queries R^+, R^- and $R^=$ for which the run "$ALL_{Selection}$" performs better (lower, equal to) than run "$STE_{NoSelection}$" in terms of informativeness metric related to the Bigrams with 2-gaps

R set	#Queries (%)	Avg. % change
R^+	27/47 (57.44%)	+12%
$R^=$	0/47 (00.00%)	–
R^-	20/47(42.55%)	−8%

This observation is confirmed by the bilateral paired *Student t-test* to evaluate the statistical significance of the average differences between these two runs. The differences, respectively for Unigrams, Bigrams or Bigrams with 2 gaps, are not significant according to a significance level (p value) equals to 0.05.

Comparing the two runs $ALL_{Selection}$ (our best run) and $STE_{NoSelection}$ (the best run at INEX 2014), we find that for three topics (with tweet ids 257798105473380352, 262290292173045762 and 276815901897146368), the run $ALL_{Selection}$ largely underperforms the run $STE_{NoSelection}$. This is mainly due to the fact that some terms used for these expansion are too general and generate noise in the results.

Then, we compare the runs $STE_{NoSelection}$ and $ALL_{Selection}$ for the topics for which $ALL_{Selection}$ does not fail, as presented in Table 15. The respective differences, for the three evaluation measures, between these runs on these 44 topics from the 47 official ones are all statistically significant (noted † in Table 15) according to bilateral paired *Student t-tests* with significance p value equal to 0.05. This result shows that there is still room for the improvement of the terms filtering.

6.3.3 Results on INEX 2013 collection

These additional tests are based on the parameter values set for INEX 2014. We want to find out if our proposals are robust against this other set of topics knowing that, as explained before, the topic tweets of INEX 2013 are very different and more difficult to tackle than those of 2014.

Table 16[16] highlights our obtained results where the lowest scores represent the best runs, ranked according to the Informativeness for bigrams with 2-gaps.

What we get from Table 16 is that:

- Our best result, with configuration $STE_{Selection}$ that filters the statistical expansion terms, is ranked above the median run. When considering valid runs of INEX 2013, this run should be ranked 9 on 21. This result is far from the best run of 2013, but we do not consider hashtags (unlike at least the top four official runs);
- In this case of difficult topics, our statistical expansion outperforms our other proposals. These findings are similar to what we obtained on the TREC 2011 microblog search collection;
- Here again, the selection proposed plays a positive role during the query expansion for all configurations;

[16] For post INEX 2013 experiments, the runs $ALL_{NoSelection}$ could not be tested because the system provided by the INEX organizers is no more accessible.

Table 15 Informativeness results on the 44 selected runs

Configuration	Unigrams	Bigrams	Bigramswith2-gaps
$ALL_{Selection}$	0.7364†	0.8439†	0.8454†
$STE_{NoSelection}$(361)	0.7800	0.8912	0.8923

Table 16 Post INEX Tweet Contextualization 2013 results, with official informativeness results Bellot et al. (2016)

Run	Configuration	Unigrams	Bigrams	Bigrams with 2-gaps
Best Run INEX 2013 (258)	/	**0.7939**	**0.8908**	**0.8943**
Median Run INEX 2013 (278)	/	0.8673	0.9540	0.9575
Worst Run INEX 2013 (269)	/	0.9981	0.9999	0.9999
/	$STE_{Selection}$	0.8259	0.9310	0.9302
/	$SE_{Selection}$	0.8172	0.9319	0.9361
/	$STE_{NoSelection}$	0.8279	0.9356	0.9362
/	$ALL_{Selection}$	0.8271	0.9374	0.9416
/	$CE_{Selection}$	0.8654	0.9478	0.9503
/	$SE_{NoSelection}$	0.8259	0.9362	0.9404
/	$CE_{NoSelection}$	0.8639	0.9524	0.9546

Bold values indicate the best results

7 Conclusion

In this paper, we propose a hybrid query expansion model (HQE) that investigates how external resources can be combined to association rules mining and used to enhance expansion terms generation and selection. The HQE model expands queries through two phases, namely, the candidate terms generation phase and the selection phase. We used, for the first phase, local, global and external methods to generate new related terms for a query. For the second phase, we proposed a measure that computes the relatedness between a query and the set of candidate terms based on Explicit Semantic Analysis and the Confidence metric.

The HQE combines local, global and external QE methods and allows the generation of different candidate terms sets for a given query and to filtrate these sets keeping only the terms most related to the query. Among the large set of experiments related to ad-hoc retrieval (on TREC Microblog search 2011, and hard topics of TREC Robust track 2004) and tweet contextualization (CLEF INEX 2013 and 2014), we found that the proposed filtering is able to enhance results in all when the linking with external data is of good quality. For retrieval, statistical approaches using association rules obtained the best results, while for tweet contextualization, the integration of several expansions is better.

In our future work, we propose to weight the query terms to add more importance to the original query terms in order to avoid any kind of query drift. Such weights may be the relatedness scores as defined in formula (16), but we may also consider the redundancy between several expansions: for instance, if the same expansion term is proposed from

statistical expansion STE and semantic expansion SE, then we are more confident in this expansion term. Furthermore, we will investigate how to enhance the proposed QE expansion model using embedding vectors, as in Almasri et al. (2016): in this case, the control of the expansion using such approaches needs to carefully filter out unadequate terms as embedding relies on similar contexts of word usage.

Acknowledgements This work is partially supported by the French-Tunisian project PHC-Utique RIMS-FD 14G 1404.

References

Aggarwal, N., & Buitelaar, P. (2012). Query expansion using wikipedia and DBpedia. In *CLEF evaluation labs and workshop, online working notes, Rome, Italy, September 17–20, 2012*, *CEUR workshop proceedings* (Vol. 1178).

Agrawal, R., & Skirant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th international conference on very large databases, VLDB 1994, Santiago, Chile* (pp. 478–499).

Agrawal, R., Imielinski, T., & Swami, A. N. (1993). Mining association rules between sets of items in large databases In *Proceedings of the 1993 ACM SIGMOD international conference on management of data, Washington, D.C., May 26–28, 1993* (pp. 207–216).

Al-Shboul, B., & Myaeng, S.-H. (2014). Wikipedia-based query phrase expansion in patent class search. *Information Retrieval*, *17*(5), 430–451.

Almasri, M., Berrut, C., & Chevallet, J. (2013). Wikipedia-based semantic query enrichment. In *ESAIR'13, proceedings of the sixth international workshop on exploiting semantic annotations in information retrieval, co-located with CIKM 2013, San Francisco, CA, USA, October 28, 2013* (pp. 5–8).

Almasri, M., Berrut, C., & Chevallet, J. (2016). A comparison of deep learning based query expansion with pseudo-relevance feedback and mutual information, in advances. In *Information retrieval—38th European conference on IR research, ECIR 2016, Padua, Italy, March 20–23, 2016, proceedings* (pp. 709–715).

Bandyopadhyay, A., Ghosh, K., Majumder, P., & Mitra, M. (2012). Query expansion for microblog retrieval. *IJWS*, *1*(4), 368–380.

Barker, K., & Cornacchia, N. (2000). Using noun phrase heads to extract document keyphrases. In *Proceedings of the 13th biennial conference of the Canadian society on computational studies of intelligence: advances in artificial intelligence, Springer, London, UK* (pp. 40–52).

Belalem, G., Abbache, A., Belkredim, F. Z., & Meziane, F. (2016). Arabic query expansion using wordnet and association rules. *International Journal of Intelligent Information Technologies*, *12*(3), 51–64.

Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., & Tannier, X. (2014). Overview of INEX tweet contextualization 2014 track. In *Working notes for CLEF 2014 conference, Sheffield, UK, September 15–18, 2014* (pp. 494–500).

Bellot, P., Moriceau, V., Mothe, J., SanJuan, E., & Tannier, X. (2016). INEX tweet contextualization task: Evaluation, results and lesson learned. *Information Processing & Management*, *52*(5), 801–819.

Bhogal, J., MacFarlane, A., & Smith, R. P. (2007). A review of ontology based query expansion. *Information Processing & Management*, *43*(4), 866–886.

Bouchoucha, A., Liu, X., & Nie, J.-Y. (2014). Integrating multiple resources for diversified query expansion. In *Advances in information retrieval: 36th European conference on IR research (ECIR 2014), Amsterdam, The Netherlands, April 13–16, 2014, Springer, Cham* (pp. 437–442).

Buckley, C., Salton, G., Allan, J., & Singhal, A. (1994). Automatic query expansion using SMART: TREC 3. In *Proceedings of the third text retrieval conference, TREC 1994, Gaithersburg, Maryland, USA, November 2–4, 1994* (pp. 69–80).

Cao, G., Nie, J., Gao, J., & Robertson, S. (2008). Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval, SIGIR 2008, Singapore, July 20–24, 2008* (pp. 243–250).

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

Chen, Z., & Lu, Y. (2010). Using text classification method in relevance feedback. In Intelligent Information & Database Systems, *Second international conference, ACIIDS, Hue City, Vietnam, March 24–26, 2010. Proceedings, Part II* (pp. 441–449).

Colace, F., Santo, M. D., Greco, L., & Napoletano, P. (2015). Improving relevance feedback-based query expansion by the use of a weighted word pairs approach. *JASIST*, *66*(11), 2223–2234.

Gabrilovich, E., & Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI 2007, proceedings of the 20th international joint conference on artificial intelligence, Hyderabad, India, January 6–12, 2007* (pp. 1606–1611).

Gan, L., & Hong, H. (2015). Improving query expansion for information retrieval using wikipedia. *International Journal of Database Theory and Application*, *8*(3), 27–40.

Gong, C. W., Cheang, L., & Hou, U. (2006). Multi-term web query expansion using WordNet. In S. Bressan, J. Küng, & R. Wagner (Eds.), *Database and expert systems applications: 17th international conference (DEXA 2006), Kraków, Poland, September 4–8, 2006, proceedings* (pp. 379–388).

Haddad, H., Chevallet, J. P., & Bruandet, M. F. (2000). Relations between terms discovered by association rules. In *Proceedings of the workshop on machine learning and textual information access in conjunction with PKDD 2000, Lyon, France*.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Record*, *29*(2), 1–12.

Han, L., & Chen, G. (2009). Hqe: A hybrid method for query expansion. *Expert Systems with Applications*, *36*(4), 7985–7991.

Ibekwe-Sanjuan, F., & SanJuan, E. (2004). Mining textual data through term variant clustering: The termwatch system. In *Computer-assisted information retrieval (Recherche d'Information et ses Applications)—RIAO 2004, 7th international conference, University of Avignon, France, April 26–28, 2004, Proceedings* (pp. 487–503).

Jabeur, L. B., Tamine, L., & Boughanem, M. (2012). Uprising microblogs: A Bayesian network retrieval model for tweet search. In *Proceedings of the ACM symposium on applied computing, SAC 2012, Riva, Trento, Italy, March 26–30, 2012* (pp. 943–948).

Järvelin, K., Kekäläinen, J., & Niemi, T. (2001). Expansiontool: Concept-based query expansion and construction. *Information Retrieval*, *4*(3), 231–255.

Klyuev, V., & Haralambous, Y. (2011). A query expansion technique using the EWC semantic relatedness measure. *Informatica*, *35*(4), 401–406.

Ko, Y., An, H., & Seo, J. (2008). Pseudo-relevance feedback and statistical query expansion for web snippet generation. *Information Processing Letters*, *109*(1), 18–22. https://doi.org/10.1016/j.ipl.2008.08.004.

Kwok, K., Grunfeld, L., Sun, H. L., & Deng, P. (2004). TREC 2004 robust track experiments using PIRCS. In *Proceedings of the thirteenth text retrieval conference (TREC 2004), Gaithersburg, Maryland, USA, November 16–19, 2004*.

Latiri, C., Haddad, H., & Hamrouni, T. (2012). Towards an effective automatic query expansion process using an association rule mining approach. *Journal of Intelligent Information Systems*, *39*(1), 209–247.

Lau, C. H., Li, Y., & Tjondronegoro, D. (2011). Microblog retrieval using topical features and query expansion. In *Proceedings of the twentieth text retrieval conference (TREC 2011), Gaithersburg, Maryland, November 15–18, 2011*.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, *38*(6), 1842–1845. https://doi.org/10.1109/18.165464.

Li, Y., Luk, R. W. P., Ho, E. K. S., & Chung, K. F. (2007). Improving weak ad-hoc queries using wikipedia as external corpus. In *SIGIR 2007: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval, Amsterdam, The Netherlands, July 23–27, 2007* (pp. 797–798).

Liu, C., Qi, R., & Liu, Q. (2013). Query expansion terms based on positive and negative association rules. In *IEEE third international conference on information science and technology (ICIST), 2013* (pp. 802–808).

Luo, J., Meng, B., Liu, M., Tu, X., & Zhang, K. (2012). Query expansion using explicit semantic analysis. In *Proceedings of the 4th international conference on internet multimedia computing and service (ICIMCS '12), ACM, New York, NY, USA* (pp. 123–126).

Lv, C., Qiang, R., Fan, F., & Yang, J. (2015). Knowledge-based query expansion in real-time microblog search. In G. Zuccon, S. Geva, H. Joho, F. Scholer, A. Sun, & P. Zhang (Eds.), *Information retrieval technology: 11th asia information retrieval societies conference (AIRS 2015), Brisbane, QLD, Australia, December 2–4, 2015, Springer, Cham* (pp. 43–55).

Macdonald, C., & Ounis, I. (2007). Expertise drift and query expansion in expert search. In *Proceedings of the sixteenth ACM conference on information and knowledge management (CIKM 2007), Lisbon, Portugal, November 6–10, 2007* (pp. 341–350).

Martín-Bautista, M. J., Sánchez, D., Chamorro-Martínez, J., Serrano, J., & Vila, M. A. (2004). Mining web documents to find additional query terms using fuzzy association rules. *Fuzzy Sets and Systems*, *148*(1), 85–104.

Massoudi, K., Tsagkias, M., de Rijke, M., & Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. In *Advances in information retrieval—33rd European conference on IR research (ECIR 2011), Dublin, Ireland, April 18–21, 2011* (pp. 362–367).

Meij, E., Weerkamp, W., & de Rijke, M. (2012). Adding semantics to microblog posts. In *Proceedings of the fifth international conference on web search and web data mining (WSDM 2012), Seattle, WA, USA, February 8–12, 2012* (pp. 563–572). https://doi.org/10.1145/2124295.2124364.

Morchid, M., Dufour, R., & Linéars, G. (2013). *LIA@inex2012: Combinaison de thèmes latents pour la contextualisation de tweets, in 13e Conférence Francophone sur l'Extraction et la Gestion des Connaissances*. France: Toulouse.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., & Johnson, D. (2005). Terrier information retrieval platform. In *Advances in information retrieval, 27th European conference on IR research (ECIR 2005), Santiago de Compostela, Spain, March 21–23, 2005* (pp. 517–519).

Ounis, I., Macdonald, C., Lin, J., & Soboroff, I. (2011). Overview of the TREC-2011 microblog track. In *Proceedings of TREC 2011*, http://trec.nist.gov/pubs/trec20/papers/MICROBLOG.OVERVIEW.pdf.

Selvaretnam, B., Belkhatir, M., & Messom, C. H. (2013). A coupled linguistics/statistical technique for query structure classification and its application to query expansion. In *10th International conference on fuzzy systems and knowledge discovery (FSKD 2013), Shenyang, China, July 23–25, 2013* (pp. 1105–1109). https://doi.org/10.1109/FSKD.2013.6816362.

Shekarpour, S., Höffner, K., Lehmann, J., & Auer, S. (2013). Keyword query expansion on linked data using linguistic and semantic features. In *2013 IEEE seventh international conference on semantic computing, Irvine, CA, USA, September 16–18, 2013* (pp. 191–197).

Smucker, M. D., Allan, J., & Carterette, B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on information and knowledge management (CIKM 2007), Lisbon, Portugal, November 6–10, 2007* (pp. 623–632).

Song, M., Song, I., Hu, X., & Allen, R. B. (2007). Integration of association rules and ontologies for semantic query expansion. *Data & Knowledge Engineering*, *63*(1), 63–75.

Tangpong, A., & Rungsawang, A. (2000). Applying association rules discovery in query expansion process. In *Proceedings of the 4th world multi-conference on systemics, cybernetics and informatics (SCI 2000), Orlando, Florida, USA*.

Voorhees, E. M. (2004). Overview of TREC 2004. In *Proceedings of the thirteenth text retrieval conference (TREC 2004), Gaithersburg, Maryland, USA, November 16–19, 2004*.

Wei, J., Bressan, S., & Ooi, B. C. (2000). Mining term association rules for automatic global query expansion: Methodology and preliminary results. In *Proceedings of the first international conference on web information systems engneering* (*WISE'00*).

Xu, J., & Roft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference, ACM Press, Zurich, Switzerland* (pp. 4–11).

Zaki, M. J., & Hsiao, C. (2002). CHARM: An efficient algorithm for closed association rule mining. In *Proceedings of the 2nd SIAM international conference on data mining (SDM 2002), Arlington, VA, USA* (pp. 457–473).

Zingla, M. A., Ettaleb, M., Latiri, C. C., & Slimani, Y. (2014). INEX2014: Tweet contextualization using association rules between terms. In *Working notes for CLEF 2014 conference, Sheffield, UK, September 15–18, 2014* (pp. 574–584).

Zingla, M. A., Latiri, C., Slimani, Y., Berrut, C., & Mulhem, P. (2016). Tweet contextualization approach based on wikipedia and DBpedia. In *CORIA 2016—Conférence en Recherche d'Informations et Applications—13th french information retrieval conference. CIFED 2016 Colloque International Francophone sur l'Ecrit et le Document, Toulouse, France, March 9–11, 2016* (pp. 545–560).