A Combination of Reduction and Expansion Approaches to Deal with Long Natural Language queries

Mohamed Ettaleb, Chiraz Latiri, Patrice Bellot

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 126 (2018) 768–777

Procedia Computer Science

www.elsevier.com/locate/procedia

22nd International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

A Combination of Reduction and Expansion Approaches to Deal with Long Natural Language queries

Mohamed ETTALEB[a], Chiraz LATIRI[b], Patrice BELLOT[b]

[a]*University of Tunis El Manar, Faculty of Sciences of Tunis, LIPAH research Laboratory, Tunis ,Tunisia*
[b]*Aix-Marseille University, CNRS, LIS UMR 7020, 13397, Marseille, France*

Abstract

Most of the queries submitted to search engines are composed of keywords but it is not enough for users to express their needs. Through verbose natural language queries, users can express complex or highly specific information needs. However, it is difficult for search engine to deal with this type of queries. Moreover, the emergence of social medias allows users to get opinions, suggestions or recommendations from other users about complex information needs. In order to increase the understandability of user needs, tasks as the CLEF Social Book Search Suggestion Track have been proposed from 2011 to 2016. The aim is to investigate techniques to support users in searching for books in catalogs of professional metadata and complementary social media. In this context, we introduce in the current paper a statical approach to deal with long verbose queries in Social Information Retrieval (SIR) by taking Social Book Search(SBS) as a study case. firstly, a morphosyntactic analysis was introduced to reduce verbose queries, the second step is based on expanding the reduced queries using association rules mining combined with Pseudo relevance feedback. Experiments on SBS 2014 and 2016 collections show significant improvement in the retrieval performance.

Keywords: Quey Expansion; association rules; Verbose Query Reduction ;Social Book Search;

1. Introduction

Discovery information is challenging due to the complexity of retrieval system required to identify human information needs, especially with the augmentation of the number of on-line portals and book catalogs. Books are considered the dominant information resource, and accordingly efforts are being made to help users find the required book(s). The exponential growing of social medias motivates the use of collective intelligence in a spirit of online collaboration. Recently, social tagging, social searches and personalized searches have become wide spread, users create their own

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000.
E-mail address: author@institute.xxx

10.1016/j.procs.2018.08.011

tags that reflect their interests and presences[19]. User-generated content to annotate web resources is playing a great role in improving Information Retrieval (IR) tasks. they may get opinions, suggestions, or recommendations from other members. Moreover, it is often difficult for users to express their information needs in a search engine query [7]. These social medias(*i.e.* LibraryThing, Reddit, etc) allow users to express complex or highly specific information needs through verbose queries. In this context, this type of queries provides detailed information which cannot be found in user-profile such as the expectations of the users and their needs. Verbose queries, expressed in natural language, are typified by their length and a complex structure which provide much more context for a more accurate understanding of users needs. Past research in information retrieval noticed that long queries can increase the retrieval performance. However, for web search queries, many researchers have noticed that search engines perform poorly on verbose queries[15]. The causes for poor performance are as follows: (*1*) High degree of query specificity, (*2*) Term redundancy or extraneous terms (lot of noise), and; (*4*) Difficulty in distinguishing between the key and complementary concepts. Thereby, it's important to find techniques that allow to deal effectively with such type of needs. The intrinsic characteristics of this type of query permit us to highlight the use of processes from the Natural Language Processing (NLP) to infer semantic aspects and thus potentially better interpret users needs. Verbose queries have received more attention in recent years[23][12]. They specified several ways of handling long queries: *i.e.*,Query Reduction to a single sub-query, Query Expansion, etc. Table 5.1 shows examples of the two techniques that we talked about.

Approach	**Initial Query**	**Modified Query**
Query Reduction to a single sub-query	i need new books i love books in series any recommendations? Vampire Fiction	new book book series vampire fiction
Query Expansion	I need new manga!. Manga and Anime Addicts	manga Anime Addicts /Ninja Vampire

Table 1. Examples of Approaches for Handling Verbose Queries.

Verbose queries can be reduced to a single sub-query which could be, for example, the most important noun phrase in the query. Query expansion (QE) is a widely used technique that attempts to raise the likelihood of a match between the query and relevant documents by adding semantically related words (called expansion terms) to a users query. The source of the expansion terms is an important issue in QE. Expansion terms may be taken from the whole target collection or from a few documents retrieved at top ranks in response to the initial query. Researchers have explored the idea of collecting expansion terms from the Web [6], Wikipedia[20], and query logs[11] of search engines, and so on. The majorities of QE Approaches have used the documents as source to expansion terms and it helps increase the performance for a verbose query. Our goal is to notice if QE can possibly improve the performance of a verbose query by using another sources like social informations(reviews, tags, etc) ? To our knowledge no study has focused on social informations as source of the expansion terms. In this paper, we propose to tackle verbose queries for social information retrieval, and to evaluate the results, we use social book search (SBS) collection as test.
In this study, we propose to combine two types of approaches. First, we propose to reduce the verbose queries by keeping only the appropriate terms. While the second approach is based in two techniques: QE approaches that are based on association rules mining in order to extend the reduced queries by adding related terms from social informations mentioned by users. The remainder of the paper is organized as follows: Section 2 discusses related work on query expansion for information retrieval. Section 3 explains the main goal of social book search as well as the test collection and topics used. Then, a detailed description of query expansion approaches is presented in Section 4. Section 5 describe our experimental results.

2. Query Expansion Approaches

To overcome the query disambiguation, query expansion plays a major role in improving the Internet searches, where the user's original query is reformulated to a new query by adding new appropriate terms with high significance. Query expansion has received a great deal of attention for resolving the short query problem. Early work on automatic query expansion dates back to the 1960s. Rocchios relevance feedback method[26] is still used in its original and modified forms for QE. On the basis of data sources used in query expansion, several approaches have been proposed. All these approaches can be classified into three main groups: (1) Global analysis, (2) Local analysis and (3) external analysis. In this section, we will mention the advances made by these approaches.

Local analysis: Local analysis includes query expansion approaches that select expansion terms from documents collection retrieved in response to the user's original query. Using local analysis, there are two ways to expand user's original query: (1) Relevance feedback(RF) and (2) Pseudo-relevance feedback(PRF). In RF, user's feedback about documents retrieved in response the original query is collected, the feedback is about whether or not retrieved documents are relevant to the user's query. The query is reformulated based on documents found relevant as per user's feedback. Rocchio's method [26] was amongst the first to use relevance feedback. This method used an information retrieval system based on the vector space model. The main idea behind this approach is to modify the user's original query vector based on the user's feedback. Another approach alike to relevance feedback approach is Pseudo-relevance feedback. This directly uses the top retrieved documents in response to user's original query for composing query expansion terms. The user is not involved here in selection of relevant documents. Rocchio's method [26] can also be applied in the context of PRF, where every particular term extracted from the top retrieved documents is assigned a score by employing a weighting function to the entire collection. This technique was first proposed in [9], which employs it in a probabilistic model. [8] 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. A recent work [27] uses fuzzy logic-based query expansion techniques and selects top-retrieved documents based on pseudo-relevance feedback. Here, each expansion term is assigned a distinct relevance score using fuzzy rules. Terms having highest scores are selected for query expansion. Moreover, a recent trend in the literature of QE is the use of word embeddings to select candidate terms [4], [13], [14]. Word embeddings are a representation of words as dense vectors in a low-dimensional vector space. These vectors estimation is based on the idea that words in similar contexts have similar meanings. Roughly speaking, Word embeddings carry relationships between terms, such as a city and the country it belongs to, *e.g.*, France is to Paris what Germany is to Berlin.

Global analysis: Query expansion techniques implicitly select expansion terms from hand-built knowledge resources for expanding/reformulating the initial query. Only individual query terms are considered for expanding the initial query and expansion terms are semantically similar to the original terms. Each term is assigned a weight and expansion terms can be assigned less weight in comparison to the original query terms. Candidate terms are usually identified by mining term-term relationships from the target corpus[24]. In [28], authors utilize the rich semantics of domain ontology and evaluates the trade off between the improvement in retrieval effectiveness and the computation cost. Several research works have been done on query expansion using a thesaurus.

In addition to the global approach based on Thesaurus construction, a promising track consists in the application of text mining methods to extract hidden and valuable dependencies between terms. One way to achieve this is by applying association rule mining in order to retrieve correlated patterns[3]. A pattern can be any set of terms and an association rule binds two sets of terms: 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 in QE could increase retrieval effectiveness[17], [1].

External analysis: External QE approaches involve methods that obtain expansion terms from other resources besides the target corpus [24]. Authors in [5], for example, used external corpora as a source for query expansion terms. Specifically, they used the Google Search API. In [18], authors proposed a Twitter retrieval framework that focuses on topical features, combined with query expansion using PRF to improve information retrieval results. As motivated before, we propose to enhance the QE approach proposed in [17] by learning to rank association rules.

3. SOCIAL BOOK SEARCH

The primary goal of Social Book Search is to facilitate easy access to and search for books that a user might be interested in, based on the query he posted. Book search and retrieval can be done traditionally using indexing methodologies and information retrieval models and tools. This is a successful and classic approach since the early years. However, many suggestion tasks at present include using a recommender system which is a more state-of-the-art enhancement to the traditional ways. It is a task to satisfy the users' requests about book recommendation from the Amazon book collection with a variety of social information. Since 2011, CLEF-INEX has provided a document collection for the SBS Track with 2.8 million books that contain XML tags from both LibraryThing(LT) and Amazon(see Fig.1). Each book has its own XML document. The XML document includes, from Amazon, professional data (*e,g*, publisher, title, creator, subject) and social data (*e,g*., reviews and tags) along with user-generated content in

the form of user reviews and ratings. LT users discuss their books on the discussion forums. Many of the topic threads are started with a request from a member for interesting, fun new books to read. Users typically describe what they are looking for, give examples of what they like and do not like, indicate which books they already know and ask other members for recommendations. Members often reply with links to works catalogued on LT, which, in turn, have direct links to the corresponding records on Amazon. To evaluate systems participating in the SBS, a set of topics have been also made available. A sample topic is presented in Figure 2. Each topic associated with a user has his/her user profile included as part of the topic. The user profile contains information about his/her interests. These interests are usually found by analyzing the catalogue of a user (which is part of his/her profile). The genre (e.g., fiction) included in his catalogue is based on that of the majority of books in the catalog. The topics are all in the mediated query which is an expanded version of the title. The narrative states what the user is looking for in a descriptive manner. As of examples field, it consists of a little number of similar books to the request that some LibraryThing users provide in their topics in order to indicate the kind of books they request. Apart from the topics and the corpus, a set of 94,000 anonymized user profiles from LibraryThing are also provided. This is used to help generate recommendations for topics (using collaborative filtering) in conjunction with a recommender system developed by the teams. Each profile contains information by the anonymous user about catalogued books. The originality of this task is to manage with verbose queries and increase the understanding of user needs. For this reason, we proposed to combine two methods to improve the understanding of the queries. Firstly, we decided to filter the initial query by keeping only the appropriate terms. Secondly, we proposed keyword expansion module based on association rules between terms.

```
<book>
<isbn>0007247001</isbn>
<title>Collins Korean Phrasebook CD Pack: The Right Word in Your Pocket (Collins Gem)</title>
<publisher>HarperCollins UK</publisher>
<publicationdate>2008-05-28</publicationdate>
<reviews>
<review>
<authorid>A22UA55AYSSN7G</authorid>
<date>2009-02-09</date>
<summary>Okay, but could be better</summary>
<content>
I am moving to Korea next month and bought this book (in a store) because it came with a CD which.
 I thought would be useful for pronounciation accuracy.
 <br /> <br />I was right about the CD, however the book is somewhat lacking.
<rating>3</rating>
<totalvotes>0</totalvotes>
<helpfulvotes>0</helpfulvotes>
</review>
</reviews>
<tags>
<tag count="1">Korean Language Dictionaries</tag>
<tag count="1">Korean Language Phrasebooks</tag>
</tags>
<similarproducts>
<similarproduct>1411669630</similarproduct>
<similarproduct>1857333659</similarproduct>
</similarproducts>
</book>
```

Fig. 1. Example of reviews for a book

4. PROPOSED APPROACH

This section illustrates the methods used for improving information retrieval and processing verbose queries before they are submitted to an information retrieval system. The methods depicted in Figure 3 that have been adopted are the following:

- stopword removal and stemming for the English language to reduce the verbose queries
- query expansion based on Associations rules between terms.

```
<topic>
  <topicid>3994</topicid>
  <request>I'm not too happy with my current thesaurus, Bartlett's Roget's Thesaurus. Does anyone have a favorite?
        One of the members of this group saw this query on the under-used Can you recommend...
        group and suggested that I post it here. It does seem like something librarians might know about.
  </request>
  <group>Librarians who LibraryThing</group>
  <title>Can you recommend a good thesaurus?</title>
  <examples/>
  <work/>
<topic>
```

Fig. 2. Example of user request(the request itself and a list of informations about his/her interests)

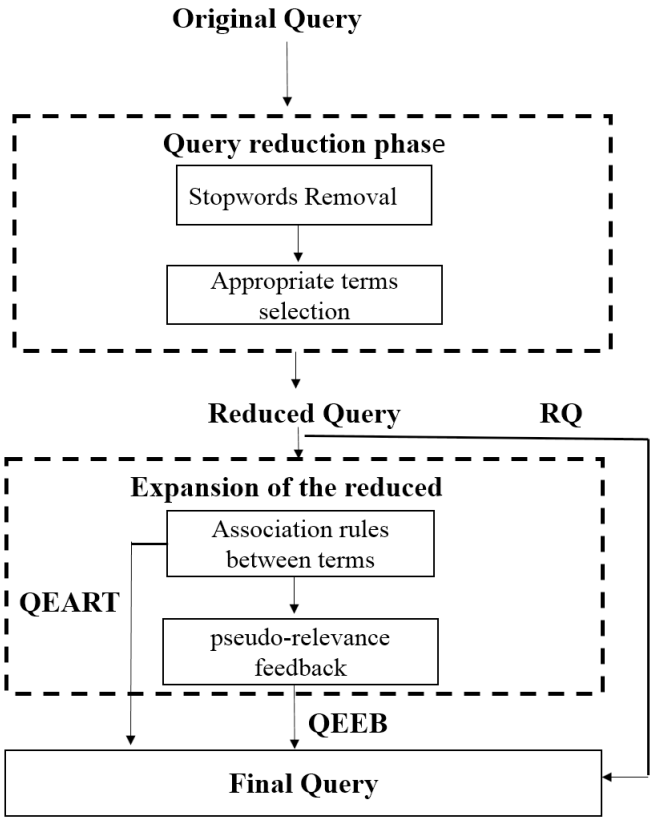

Fig. 3. The proposed Reduction and Expansion Approaches

- query expansion using similar books mentioned in the topics.

4.1. Query Reduction

Removing stopwords has long been a standard query processing step [10]. We used three different stopwords lists in this study: the standard stopword list[1], as well as two stopwords lists based on morphosyntactic analysis and according to the ranks of terms by some weight. [21] present several methods for automatically constructing a collection dependent stopword list. Their methods generally involve ranking collection terms by some weight, then

[1] https://www.ranks.nl/stopwords

choosing some rank threshold above which terms are considered stopwords. [16] constructed a specific stopword list of a given collection and used statistic measure Inverse Document Frequency(IDF) to rank terms and decide which term is a stopword or not. Next, they applied these techniques by removing from the query all words which occur on the stopword list. Our proposal is to reduce the verbose queries based on two steps: first, all terms that appear in the standard stopwords list are eliminated. Second, we process the linguistic filtering method and execute TreeTagger[2] *a part of speech tagging* on the queries. Then, we select only the particular words of *noun type*(nouns, proper nouns, etc.) and query words that have a form as noun phrase such as Syrian Civil War. The aim is to keep the appropriate words that can improve the quality of the user query.

4.2. Query expansion using Associations rules between terms(ART)

Query expansion is the process of adding additional relevant terms to the original queries to improve the performance of information retrieval systems. However, previous studies showed that automatic query expansion using Association rules do not lead to an improvement in the performance.

The main idea of this approach is to extract a set of non redundant rules, representing inter-terms correlations in a contextual manner. We use the rules that convey the most interesting correlations amongst terms, to extend the initial queries. Then, we extract a list of books for each query using the MB25 scoring [25].

4.2.1. Representation and Query Expansion:

We represent a query q as a bag of terms:

$$q = \{t_1, ..., t_n\} \tag{1}$$

where t_i is a term in q.
Given an original query q_i. the process for obtaining the associated expanded query denoted eq is consisting to select a set of candidate terms for q:

$$CT = \{w_1, ..., w_m\} \tag{2}$$

Where w_i is a candidate term. This set of candidate terms, denoted CT, is selected using association rules between terms detailed in the next section.

4.2.2. Association Rules:

An association rule, *i.e.,* between terms, is an implication of the form $R : T1 \Rightarrow T2$, where $T1$ and $T2$ are subsets of τ where $\tau := \{t_1, ..., w_l\}$ is a finite set of l distinct terms in the books collection and $T1 \cap T2 = \emptyset$. The termsets $T1$ and $T2$ are, respectively, called the *premise* and the *conclusion* of R. The rule R is said to be based on the termset T equal to $T1 \cup T2$. The *support* of a rule $R : T1 \Rightarrow T2$ is then defined as:

$$Supp(R) = Supp(T) \tag{3}$$

while its confidence is computed as:

$$Conf(R) = \frac{Supp(T)}{Supp(T_1)} \tag{4}$$

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.

4.2.3. Candidate Terms Generation Approach based on Association Rules:

The main idea of this approach is to use the association rules mining technique to discover strong correlations between terms [2]. The set of query terms will be expanded using the maximal possible set of terms located in the conclusion parts of the retained rules while checking that the terms are located in their premise part. An illustrative example of association rules is highlighted in Table4.2.3.

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

R	Support	Confidence
military ⇒ warfare	83	0.741
romance ⇒ love	64	0.723

Table 2. Association Rules examples.

The process of generating candidate terms for a given query is performed as in the following steps:
1) Selection of a sub-set of 12000 books according to the querys subject. The books are represented only by their social information, we chose to select the title, reviews and the tags as content of the book.
2) Annotating the selected books using TreeTagger. The choice of TreeTagger was based on the ability of this tool to recognize the nature(morphosyntactic category) of a word in its context.
3) Extraction of nouns (terms) from the annotated books, and removing the most frequent ones.
4) Generating the association rules using an efficient algorithm: Closed Association Rule Mining (CHARM)[29] for mining all the closed frequent termsets, As parameters, CHARM takes minsupp = 15 as the relative minimal support and minconf = 0.7 as the minimum confidence of the association rules [17]. While considering the Zipf distribution of the collection, the maximum threshold of the support values is experimentally set in order to spread trivial terms which occur in the most of the 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 rule. CHARM gives as output, the association rules with their appropriate support and confidence. Table4.2.3 describes the output of CHARM.

4.3. Query expansion using pseudo-relevance feedback

The query in pseudo-relevance feedback technique is submitted to the search engine and the top results are extracted and considered as being relevant to the query. These related documents are then scanned for more keywords related to the query. In our case, we assumed that the list of similar books(Figure 2) mentioned by users in their topics are relevant and contain terms can be important to the query. In order to exploit these similar books, we expand again the queries processed in the section 4.2 by automatically adding terms from these similar books.

5. Experiments and results

In this section, we describe the experimental setup we used for our experiments.

5.1. Experimental data

To evaluate our approach, the data provided by CLEF SBS suggestion track 2016[3] are used.

- Documents: The documents collection consists of 2.8 millions of books descriptions with meta-data from Amazon and LibraryThing. Each document is represented by book-title, author, publisher, publication year, library classification codes and user-generated content in the form of user ratings and reviews.
- Queries: the collection of queries from 2011 to 2016 the organizers of SBS have used Librarything forum to extract a different set of queries with relevance judgments for each year. In our case, we chose to combine the title with the narrative as a representation of the queries.

For fair comparison, the queries and the corresponding relevance judgments in the others years are utilized as the training set when evaluating on the each-year dataset, which is the same to the 6-year SBS evaluation campaigns. As for the evaluation metrics, NDCG@10, P@10 and MAP are chosen and the official results are sorted based on the scores of NDCG@10, according to the official organizer of SBS.

5.2. Experimental setup

In our experiments, we present experimental results on SBS 2014 and 2016 dataset with Amazon collections to compare the performances of different components of our system. First, we used *Terrier Information Retrieval* framework developed at the University of Glasgow [22]. Terrier is a modular platform for rapid development of large-scale information retrieval applications. It provides indexing and retrieval functionalities. The BM25 model was used

[3] http://social-book-search.humanities.uva.nl/#/data/suggestion

Year	#Queries	Fields
2011	211	Title,Group,Narrative,type,genre,specificity
2012	96	Title,Group,Narrative,type,genre
2013	370	Title,Group,Narrative,Query
2014	672	Title,Group,Narrative,mediated_query
2015	178	Title,Group,Narrative,mediated_query
2016	119	Title,Group,Request

Table 3. The six years topics used for SBS Suggestion track

for querying with the usual parameter values ($b = 0, k3 = 1000, k1 = 2$). Using the BM25 model, the relevance score of a book d for query Q is given by:

$$S(D,Q) = \sum_{t \in Q} \frac{(K_1+1)w(t,d)}{K_1 + w(t,d)} . idf(t) . \frac{(K_3+1)w(t,Q)}{K_3 + w(t,Q)} \tag{5}$$

Where $w(t,d)$ and $w(t,Q)$ are respectively the weights of terms in document D and in query Q. $idf(t)$ is the inverse document frequency of term t, given as follow:

$$idf(t) = log \frac{|D| - df(t) + 0.5}{df(t) + 0.5} \tag{6}$$

Where $df(t)$ is the number of documents containing t, and $|D|$ is the number of documents in the collection. We conducted three different runs, namely:

1. Run-**RQ:** We used only the reduced queries we showed in the section4.1.
2. Run-**QEART:** We added the association rules between terms to extend the reduced queries.
3. Run-**QEEB:** Query expansion using examples books.

Strategy	NDCG@10	MAP	Improved
Baseline model	0.1041	0.0965	-
RQ	0.1158	0.1014	11.24%
QEART	0.1429	0.1153	**23.4%**
QEEB	**0.1518**	**0.1194**	6.23%

Table 4. Results of SBS 2014 with different strategies

Strategy	NDCG@10	MAP	Improved
Baseline model	0.1175	0.0872	-
RQ	0.1240	0.0904	5.53%
QEART	0.1549	0.1013	**24.92%**
QEEB	**0.1688**	**0.1054**	8.97%

Table 5. Results of SBS 2016 with different strategies

We used two topic sets provided by CLEF SBS in 2014 (680 topics) and 2016 (120 topic). We selected the title and narrative fields for each topic. In the beginning, we used the techniques we showed in the section4.1 to remove the stop-words and keep the appropriate words in the query. Secondly, the reduced query was expanded by adding new terms using **ART**. In this step, we applied the CHARM algorithm with the following parameters : $minsup = 15$, and $minconf = 0.7$. Then, we used the similar books for each topic and applied the pseudo-relevance feedback technique to expand again the query. The rocchio function was used with its default parameter settings $\beta = 0.4$, and the number of terms selected from each similar book was set to 10. Table5.2 describe an example of both approaches based on reduced query and query expansion based on association rules between terms.

Original Query	Does anyone know of a good book on the Battle of Gazala?
Reduced Query	good book battle gazala
Query Expansion using ART	good book battle gazala / military history gazala war attack army

Table 6. Examples of reduced and expansion Approaches for handling verbose queries

5.3. Experimental results

We first compare our baseline retrieval results with results from different expansion strategies which are shown in Table5.2 and Table5.2. The columns RQ, QEART and QEEB represent the results obtained by the reduced query, expanded reduced query using association rules and expanded (QEART) using examples books respectively. As shown

in the two tables, with the proposed different expansion strategies, the results are well-performed and improve the baseline to some extent. We notice that when we used the query reduction technique, the results perform better than the baseline in the two sets of topics. we also notice that when applying query expansion technique using pseudo-relevance feedback, the results are better across all the sets of topics. In term of ndcg@10 the results increase from 0.1429 to 0.1518 in the set of 2014 and from 0.1549 to 0.1688 in the set of 2016. From the overall perspective, the best performance is obtained by QEART strategy with the greatest improvements of the score of 24.9% in the score of NDCG@10. This can be explained by the fact that the added terms are chosen based on the confidence of the association rule where we consider only the rules of high confidence, *i.e.,* using the strongest correlations inter-terms lead to generate precise terms. Tables 5.3 and 5.3 present the comparative results. Our run is best-performed in the evaluation on 2014 datasets, and our best results in 2016 are ranked second in term of NDCG@10. As to map, we can notice that our approach outperforms the best runs in the year 2014 but is still lower than the best runs of the year 2016. Concerning the best run that wins the 2016 SBS campaign (Official run), the authors proposed a searching framework which builds at any moment a reading list for any specific topic, where the relevance between topics and books, the books quality, the popularities timeliness and the results diversity are respectively embedded into vector representations based on user-generated contents and statistics on social media. The obtained evaluation results also

Run	NDCG@10	MAP
Our run	**0.1518**	**0.1194**
Official run	0.1420	0.102
Medium run	0.096	0.068
Worst run	0.010	0.007

Table 7. Comparison results on Social Book Search 2014.

Run	NDCG@10	MAP
Our run	0.1688	0.1054
Official run	**0.2157**	**0.1253**
Medium run	0.0861	0.0524
Worst run	0.0018	0.0004

Table 8. Comparison results on Social Book Search 2016.

shed light on the fact that our proposed approaches offer interesting results. However, we noticed that the QEART worked well in the reviews also, this is justified by the fact that the association rules allowed us to find the terms having a strong correlation with the querys terms.

Finally, to further the effectiveness analysis, we present a gain and failure analysis on our approach. Table 5.3 presents the percentages of queries R^+ and R^- for which QE techniques perform better or lower/equal than the different baselines in terms of NDCG@10. As depicted in Table 5.3, the average percentage for the set of queries R^+ is of about 67.40% for the SBS 2014 collection and 66.11% for SBS 2016 collection. The high percentage for R+ queries is reached when we combined the ART technique with PRF for QE. these results confirm the effectiveness of using mainly the association rules as well as PRF in query expansion as proved in the literature.

Run	QEART	QEEB
SBS 2014 COLLECTIONS		
R^+	62.55	72.24
R^-	37.45	29.16
SBS 2016 COLLECTIONS		
R^+	61.39	70.83
R^-	38.61	29.17

Table 9. Percentage of queries R+ and R for each set of query (better or lower/equal) than the different baselines in terms of NDCG.

6. Conclusion

This paper has described new methods to handling with verbose queries, we found that the nature of the queries poses a great challenge to an effective use of query expansion approaches in this context. Firstly, we presented a method to generate a stopword list in order to reduce verbose queries. Secondly, novel query expansion approaches was proposed based on association rules between sets of terms using only the social informations(tags, reviews). The obtained results confirmed that the synergy between association rules and query expansion is very fruitful. In our future work, we propose to weight the terms of verbose queries to add more importance on the original query terms. We also propose to use other aspects to define which queries are better suited for query reduction and which queries can be better improved through query expansion.

References

[1] Abbache, A., Meziane, F., Belalem, G., Belkredim, F.Z., 2016. Arabic query expansion using wordnet and association rules. IJIIT 12, 51–64. URL: https://doi.org/10.4018/IJIIT.2016070104, doi:10.4018/IJIIT.2016070104.

[2] Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, New York, NY, USA. pp. 207–216.

[3] Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th International Conference on Very Large Data Bases, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 487–499.

[4] Almasri, M., Berrut, C., Chevallet, J.P., 2016. A Comparison of Deep Learning Based Query Expansion with Pseudo-Relevance Feedback and Mutual Information, in: Conférence ECIR, Padoue, Italy. pp. 369 – 715. URL: https://hal.archives-ouvertes.fr/hal-01576603, doi:10.1007/978-3-319-30671-1_57.

[5] Bandyopadhyay, A., Ghosh, K., Majumder, P., Mitra, M., 2012. Query expansion for microblog retrieval. IJWS 1, 368–380. URL: https://doi.org/10.1504/IJWS.2012.052535, doi:10.1504/IJWS.2012.052535.

[6] Bendersky, M., Metzler, D., Croft, W.B., 2012. Effective query formulation with multiple information sources, in: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, ACM, New York, NY, USA. pp. 443–452.

[7] Biskri, I., Rompre, L., 2012. Using associated rules for query reformulation, in: Next Generation Search Engine: Advanced Models for Information Retrieval. IGI-Global, pp. 291–303.

[8] Cao, G., Nie, J.Y., 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, ACM, New York, NY, USA. pp. 243–250. URL: http://doi.acm.org/10.1145/1390334.1390377, doi:10.1145/1390334.1390377.

[9] Croft, W.B., Harper, D.J., 1979. Using probabilistic models of document retrieval without relevance information. Journal of Documentation 35, 285–295. URL: https://doi.org/10.1108/eb026683, doi:10.1108/eb026683.

[10] Croft, W.B., Metzler, D., Strohman, T., 2009. Search Engines - Information Retrieval in Practice. Pearson Education. URL: http://www.search-engines-book.com/.

[11] Cui, H., Wen, J.R., Nie, J.Y., Ma, W.Y., 2002. Probabilistic query expansion using query logs, in: Proceedings of the 11th International Conference on World Wide Web, ACM, New York, NY, USA. pp. 325–332.

[12] Cummins, R., 2016. A study of retrieval models for long documents and queries in information retrieval, in: Proceedings of the 25th International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland. pp. 795–805. URL: https://doi.org/10.1145/2872427.2883009, doi:10.1145/2872427.2883009.

[13] Diaz, F., Mitra, B., Craswell, N., 2016. Query expansion with locally-trained word embeddings. CoRR abs/1605.07891. URL: http://arxiv.org/abs/1605.07891, arXiv:1605.07891.

[14] Fernández-Reyes, F.C., Valadez, J.H., Montes-y-Gómez, M., 2018. A prospect-guided global query expansion strategy using word embeddings. Inf. Process. Manage. 54, 1–13. URL: https://doi.org/10.1016/j.ipm.2017.09.001, doi:10.1016/j.ipm.2017.09.001.

[15] Gupta, M., Bendersky, M., 2015. Information retrieval with verbose queries. Foundations and Trends in Information Retrieval 9, 91–208.

[16] Huston, S., Croft, W.B., 2010. Evaluating verbose query processing techniques, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA. pp. 291–298.

[17] Latiri, C.C., Haddad, H., Hamrouni, T., 2012. Towards an effective automatic query expansion process using an association rule mining approach. J. Intell. Inf. Syst. 39, 209–247. URL: https://doi.org/10.1007/s10844-011-0189-9, doi:10.1007/s10844-011-0189-9.

[18] Lau, C.H., Li, Y., Tjondronegoro, D., . Microblog retrieval using topical features and query expansion.

[19] Lee, K.P., Kim, H.G., Kim, H.J., 2012. A social inverted index for social-tagging-based information retrieval. J. Inf. Sci. 38, 313–332. URL: http://dx.doi.org/10.1177/0165551512438357, doi:10.1177/0165551512438357.

[20] Li, Y., Luk, W.P.R., Ho, K.S.E., Chung, F.L.K., 2007. Improving weak ad-hoc queries using wikipedia asexternal corpus, in: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA. pp. 797–798. URL: http://doi.acm.org/10.1145/1277741.1277914, doi:10.1145/1277741.1277914.

[21] Lo, R.T., He, B., Ounis, I., 2005. Automatically building a stopword list for an information retrieval system. JDIM 3, 3–8. URL: http://www.dirf.org/jdim/abstractv3i1.htm#01.

[22] Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C., 2006. Terrier: A high performance and scalable information retrieval platform. Proceedings of the OSIR Workshop , 18–25.

[23] Paik, J.H., Oard, D.W., 2014. A fixed-point method for weighting terms in verbose informational queries, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, New York, NY, USA. pp. 131–140.

[24] Pal, D., Mitra, M., Bhattacharya, S., 2015. Exploring query categorisation for query expansion: A study. CoRR abs/1509.05567. URL: http://arxiv.org/abs/1509.05567, arXiv:1509.05567.

[25] Robertson, S.E., Zaragoza, H., 2009. The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval 3, 333–389.

[26] Salton, G., 1971. The SMART Retrieval System—Experiments in Automatic Document Processing. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[27] Singh, J., Sharan, A., 2017. A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach. Neural Comput. Appl. 28, 2557–2580.

[28] Wu, J., Ilyas, I., Weddell, G., . A study of ontology-based query expansion.

[29] Zaki, M.J., Hsiao, C., 2002. CHARM: an efficient algorithm for closed itemset mining, in: Proceedings of the Second SIAM International Conference on Data Mining, Arlington, VA, USA, April 11-13, 2002, pp. 457–473.