Learning Query Expansion from Association Rules Between Terms

Ahem Bouziri[1], Chiraz Latiri[2], Eric Gaussier[3] and Yassin Belhareth[4]
[1]*ENSI - Manouba University, LIPAH-FST, Manouba, Tunisia*
[2]*ISAMM - Manouba University, LIPAH-FST, Tunis, Tunisia*
[3]*Université Joseph Fourrier-Laboratoire d'Informatique de Grenoble, Grenoble, France*
[4]*ISAMM - Manouba University, Manouba, Tunisia*

Keywords: Query Expansion, Association Rules, Classification.

Abstract: Query expansion technique offers an interesting solution for obtaining a complete answer to a user query while preserving the quality of retained documents. This mainly relies on an accurate choice of the added terms to an initial query. In this paper, we attempt to use data mining methods to extract dependencies between terms, namely a generic basis of association rules between terms. Face to the huge number of derived association rules and in order to select the optimal combination of query terms from the generic basis, we propose to model the problem as a classification problem and solve it using a supervised learning algorithm. For this purpose, we first generate a training set using a genetic algorithm based approach that explores the association rules space in order to find an optimal set of expansion terms, improving the MAP of the search results, we then build a model able to predict which association rules are to be used when expanding a query. The experiments were performed on SDA 95 collection, a data collection for information retrieval. The main observation is that the hybridization of textmining techniques and query expansion in an intelligent way allows us to incorporate the good features of all of them. As this is a preliminary attempt in this direction, there is a large scope for enhancing the proposed method.

1 INTRODUCTION

Query expansion technique aims to reducing the usual query/document mismatch by expanding the query using terms that are related to the original query terms, but have not been explicitly mentioned by the user. The goal of this technique is not only to improve the recall by retrieving relevant documents that cannot be retrieved by the user query, but also to improve the precision of the retrieved documents by putting the most relevant ones at the top list of the retrieved documents. We clame that a synergy between classical IR techniques and some advanced text mining methods, especially association rules between terms (Agrawal and Skirant, 1994) is particularly appropriate.

However, applying association rules in the context of IR is far from being a trivial task, mostly because of the huge number of potentially interesting rules that can be drawn from a document collection. We mainly concentrate in this work on reducing the number of selected rules for the expansion process while retaining the most interesting ones.

In this paper, we propose to model query expansion based on association rules as a classification problem and solve it using a supervised learning algorithm. Given a query and the set of its association rules, the classifier must be able to decide for each association rule whether it is appropriate to use it for expansion. Our automatic query expansion process is based on the new generic basis of association rules. The main thrust in the proposal is that the introduced basis gathers a minimal set of rules allowing an effective selection of rules to be used in the expansion process.

2 RELATED WORK

Different methods dedicated to query expansion have been proposed in the literature such as those based on user relevance feedback (Ruthven and Lalmas, 2003), pseudo relevance feedback (Buckley et al., 1994; Mitra et al., 1998), and terms co-occurrences (Lin et al., 2008; Rungsawang et al., 1999). A recent survey of automatic query expansion approaches is proposed in (Carpineto and Romano, 2012).

Bouziri, A., Latiri, C., Gaussier, E. and Belhareth, Y..
Learning Query Expansion from Association Rules Between Terms.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR*, pages 525-530
ISBN: 978-989-758-158-8

In a user relevance feedback context, related terms come from user identified relevant documents or queries. In (Fonseca et al., 2005), Fonseca *et al.* segmented query sessions in search engine query logs into subsessions and then used association rules to extract related queries from those subsessions. They calculated the relatedness between all queries using the association rule mining model and then built a query relation graph. The query relation graph was used for identifying terms related to a given user input query. The apporoach presented in (Boughanem and Tamine, 2000) combines relevance feedback and genetic algorithms query to optimise query reformulation.

On the other hand, the pseudo relevance feedback expanded terms which come from the top *k* retrieved documents assumed to be relevant without any intervention from the user. The authors in (Buckley et al., 1994; Mitra et al., 1998) proposed approaches for expanding search engine queries. The related terms are extracted from the top documents that are returned in response to the original query using statistical heuristics, and the query is expanded using these extracted terms. The results of this approach on large collections are sometimes even negative since the assumed relevant documents retrieved by an information retrieval system are unfortunately not all relevant (Buckley et al., 1994). Another limitation of this technique is that it is using a local query expansion technique based on a set of documents retrieved for the query. As a consequence, they are more focused on the given query than global analysis. Indeed, in (Xu and Croft, 1996), the authors showed that using global analysis techniques produces results that are both more effective and more predictable than simple local feedback. In (Chifu and Mothe, 2014), the pseudo relevance feedback is used in a selective query expansion approach. To overcome some of the limits of PRF, the authors propose to use it only for difficult queries. An evolutionary approach for improving efficiency of pseudo-relevance feedback-based query expansion is proposed in (Pragati Bhatnagar, 2015). In this method, the candidate terms for query expansion are selected from an initially retrieved list of documents, ranked on the basis of co-occurrence measure of the terms with the query terms.

Association rules techniques extract relationships based on term co-occurrences where the window size used is a document. The authors of (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 association rules. Using a lower confidence threshold (10%), authors performed better results (Rungsawang et al., 1999). The same approach is proposed by (Haddad et al., 2000) performing improvement when using the APRIORI algorithm to extract association rules. The best improvements were performed with low confidence values. The main limitation of this approach consists in the huge number of generated association rules while a large part of them are redundant in the sense that several rules convey the same information. The removal of redundancy within mined rules is then a key step for improving the quality of the expansion as performed in the approach we propose in this work. A more adapted mining algorithm to text that avoids redundancy is proposed by (Latiri et al., 2012). A generic basis $\mathcal{MGB}$ of non redundant association rules between terms is first derived from the tested document collection. This compact basis is then used to blindly expand the user query considering all terms that appear in the conclusions of the irredundant association rules whose premise is contained by the original query. Experimental evaluation of this approach shows an improvement of the mean precision for the tested document collections. In the present work, we refine this approach

3 MINING ASSOCIATION RULES BETWEEN TERMS

In this work, we shall use in text mining field, the theoretical framework of Formal Concept Analysis (FCA) presented in (Ganter and Wille, 1999). First, we formalize an extraction context made up of documents and index terms, called *textual context*.

Definition 1. *A textual context is a triplet* $\mathfrak{M} := (\mathcal{C}, \mathcal{T}, I)$ *where:*

- $\mathcal{C} := \{d_1, d_2, \ldots, d_n\}$ *is a finite set of n documents of a collection.*
- $\mathcal{T} := \{t_1, t_2, \cdots, t_m\}$ *is a finite set of m distinct terms in the collection. The set* $\mathcal{T}$ *then gathers without duplication the terms of the different documents which constitute the collection.*
- $I \subseteq \mathcal{C} \times \mathcal{T}$ *is a binary* (*incidence*) *relation. Each couple* $(d, t) \in I$ *indicates that the document* $d \in \mathcal{C}$ *has the term* $t \in \mathcal{T}$.

A termset is a set of terms. The support of a termset is defined as follows.

Definition 2. *Let* $T \subseteq \mathcal{T}$. *The support of* T *in* $\mathfrak{M}$ *is equal to the number of documents in* $\mathcal{C}$ *containing all the term of* T. *The support is formally defined as follows :*

$$Supp(T) = |\{d | d \in \mathcal{C} \wedge \forall t \in T : (d, t) \in I\}| \quad (1)$$

Supp(T) *is called the absolute support of* T *in* $\mathfrak{M}$. *The relative support* (*aka frequency*) *of* $T \in \mathfrak{M}$ *is equal to* $\frac{Supp(T)}{|\mathcal{C}|}$.

A termset is said *frequent* (*aka large* or *covering*) if its terms co-occur in the collection a number of times greater than or equal to a user-defined support threshold, denoted *minsupp*. Otherwise, it is said *unfrequent* (*aka rare*).

The derivation of association rules between terms is achieved starting from the set of frequent termsets extracted from a context $\mathfrak{M}$. Many representations of frequent termsets were proposed in the literature where terms are characterized by the frequency of their co-occurrence. The ones based on *closed termsets* (Pasquier et al., 2005) and *minimal generators* (Bastide et al., 2000) are at the core of the definitions of almost all generic bases of the literature. They result from the mathematical background of FCA (Ganter and Wille, 1999), described in the next subsection.

Definition 3. *An association rule R is an implication of the form R:* $T_1 \Rightarrow T_2$, *where* T_1 *and* T_2 *are subsets of* $\mathcal{T}$, *and* $T_1 \cap T_2 = \emptyset$. *The termsets* T_1 *and* T_2 *are, respectively, called the premise and the conclusion of R. The rule R is said to be based on the termset T equal to* $T_1 \cup T_2$. *The support of a rule R:* $T_1 \Rightarrow T_2$ *is then defined as:*

$$Supp(R) = Supp(T), \tag{2}$$

while its confidence is computed as:

$$Conf(R) = \frac{Supp(T)}{Supp(T_1)}. \tag{3}$$

An association 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. Also, the given support threshold *minsupp* is used to remove rules based on termsets T that do not occur often enough, *i.e.*, rules having $Supp(T) < minsupp$.

4 LEARNING QUERY EXPANSION FROM ASSOCIATION RULES

4.1 Model and Problem Statement

In our approach, the query expansion problem is defined as follows : given an original query $OQ := \{t_1, t_2, \ldots, t_n\}$, and the set of its related association rules AR_Q ; select the association rules whose conclusion terms are the more adapted to expand OQ and return documents that meet the user need.

We model this problem as a supervised classification problem in which we attempt to predict whether to use or not a given association rule to expand the query at hand. Each observation x lie in an input space $\mathcal{X} \subseteq \mathbb{R}^d$ and is associated with a class $y \in \mathcal{Y}$, where $|\mathcal{Y}| = 2$. In the context of query expansion using association rules, $x^i \in \mathcal{X}$ denotes the vector representation of a pair query/association rule and its class $y^i \in \mathcal{Y}$ represents the class associated with x^i. y^i is belongs to positive class (ie $y^i = 1$) if in the pair query/association rule represented by x^i, the association rule is used to expand the query, it belongs to the negative class ($y^i = 0$) otherwise.

4.2 Input Space Representation

Observations are vectors in the input space $\mathcal{X} \subseteq \mathbb{R}^d (d = 14)$. Values on these vectors are computed based on statistical distribution measures of terms in the query's text and in association rule conclusion. These features are of three categories and are explained in the following paragraphs.

4.2.1 Document Frequency based Features

The document frequency (DF) is a statistical predictor that measures whether a term is rare or common in the corpus. Its value for a query represents the average of the DF for all query terms . The $DF(Q)$ of a query Q is computed as in 4.

$$DF(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{|\{d_j, t_i \in d_j\}|}{|D|} \tag{4}$$

We compute also the inverted document frequency (iDF) for a query as follows :

$$iDF(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \log \frac{|D|}{|\{d_j, t_i \in d_j\}|} \tag{5}$$

Both DF and iDF are also computed for an association rule, they represent the average of the DF or iDF for all the terms in the conclusion of the association rule.

4.2.2 Term Frequency based Features

We include features dealing with term frequency in the documents of the collection. The term frequency ($TF(t,d)$) is simply the number of occurrences of term t in document d. For each term of the query or of the association rule conclusion we use an average of its frequencies in all documents and we note it

$ATF(t_i)$ for term t_i. For a query, the term frequency is considered to be the average of the ATF of al the terms in the query and is calculated as in 6.

$$TF(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|} ATF(t_i) \quad (6)$$

$$ATF(t_i) = \frac{1}{|D|}\sum_{j=1}^{|D|} TF(t, d_j)$$

When dealing with an average, it is important to know how the values are distributed around it. We introduce a variance measure to evaluate the variation of term frequencies. For a query, we compute the average of theses variations for all query terms as mentionned in equation 7

$$V(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|} \frac{1}{|D|-1}\sum_{j=1}^{|D|} (TF(t_i, d_j) - ATF(t_i, d_j))^2 \quad (7)$$

4.2.3 Features Based on Association Rule Properties

In addition to the features based on term's distribution measures, we use 6 features to characterise an association rule :

- number of terms in the association rule premise,
- number of terms in the association rule conclusion,
- association rule confidence given in (3),
- association rule support as given in(2),
- a ratio of the number of terms in the association rule premise by the number of terms in the query. This ratio measures how much the rule is matching the query.
- the proportion of terms in the association rule conclusion that are present in the other association rules. This proportion measures the importance of the conclusion terms.

4.3 Generating Training Instances

We represent training instances as couples (query / association rule). For each query, we generate as many instances as association rules with at least one term of the premise is in the query. An instance belongs to positive class if the corresponding association rule is used to find the optimal expanded query, it belongs to negative class otherwise. Query expansion is optimised using an exploring process based on genetic algorithms. The original query and the set of associations rules candidate to expansion are the input to the optimising process. The process returns an optimal expanded query and generates training instances corresponding to this optimal expansion. The principles of this exploring process are described in the following paragraphs.

4.3.1 Chromosome Representation

A chromosome is a vector of terms representing a candidate expanded query formed by the initial query terms and the conclusion terms of a candidate association rule. An association rule is considered to be candidate if its premise is composed of a sub set of initial query terms. The length of a chromosome is determined by the number of initial query added to the total number of terms candidate to expansion.

Given an original query OQ = $\{t_1,\ldots,t_n\}$, the set of terms candidate to expansion obtained from association rules of the $\mathcal{MGB}$ basis , denoted TC, is : (Latiri et al., 2012) :

$$\forall R : T_1 \Rightarrow T_2, \text{a non-redondant rule} \in \mathcal{MGB}; \quad \text{if } T_1 \subseteq \text{OQ then } TC = TC \cup T_2. \quad (8)$$

Equation (8) means if the premise of association rule R is contained inOQ, conclusion terms of R are candidate to expansion.

We adopt a binary coding of chromosomes which indicates whether a given term appear or not in the expanded query.

4.3.2 Initial Population

To build initial population individuals we first filter the generic basis of association rules $\mathcal{MGB}$ keeping only rules which premise terms are terms of the initial query. The size of initial population is equal to the number of these rules.

4.3.3 Fitness Function

Query expansion aims at improving search results. The fitness of an individual measures the ability of the candidate expanded query to return documents meeting the user's need. We use the mean average precision (MAP) returned by the SRI Terrier to measure the fitness of a chromosome.

4.3.4 Genetic Operators

Selection. This operator allows individuals in the current generation to survive, to reproduce or die. In general, the probability of survival of an individual

depends on its fitness. In this implementation , we opt for an elite selection by selecting the best *n* individuals needed to build new generation.

Crossover. The crossover operator allows recombination of the information in the genetic heritage thereby favouring exploration of the search space . The crossing permits in our case to produce new expanded query. We opt for a crossover operator which produces one child from two parents. The individual child inherits the terms of expanding both parents.

Replacement. At each iteration, new individuals replace their parents.

5 EXPERIMENTAL EVALUATION

5.1 Evaluation Framework

Experiments were conducted under TERRIER using the collection of French texts of CLEF 2003 corpus. The collection French SDA 95, noted in the remainder of this paper as SDA-95, includes 42615 documents and 60 queries each being provided with a set of relevant documents. We configure Terrier so as search includes terms in both description and narration fields.

5.2 Evaluation Steps

The evaluation of the automatic query expansion process proposed in this paper proceeds in eight steps.

1. Evaluate search with original queries to determine a baseline for comparaison.
2. Evaluate search with pseudo relevance feed back (PRF) expanded query.
3. Generate association rules minimal basis using Charm tool
4. Build trainning set : this task is automatically performed.
5. Construct learning models using decision tree (DT) under Weka.
6. Apply classifier to test set
7. Expand queries in the test set according to classification results obtained in step 3.
8. Carry out search using the weighting schema OKAPI BM25 (Jones et al., 2000).

   The relevance of the returned documents is estimated according to the following measures:

- The MAP (Mean Average Precision) which defines the overall performance of the search engine.
- Precisions at P@5, P@10, P@15, and P@30 returned relevant documents.
- Precisions at 11 recall points (P@11).

5.3 Preliminary Results and Discussion

Table 1: Experimental Results for SDA-95 Collection.

	MAP	P@5	P@10	P@20
AG	0,422	0,469	0,360	0,243
DT	0,345	0,376	0,282	0,203
PRF	0,356	0,385	0,300	0,217
Baseline	0,339	0,369	0,283	0,205
Δ AG	24,48%	27%	27%	19%
Δ DT	1,77%	1%	0%	0%
Δ PRF	5,01%	4%	6%	6%

Table 1 synthesises results obtained for the SDA95 collection . Values in AG line are obtained with expanded queries optimised through GA based exploring algorithm. These results can be considered as upperbounds for expanded queries using association rules. Values on the DT line are drawn from experiements on the same queries expanded using association rules that are predicted through classifier Weka *J*48 implementing decision tree. PRF line reports results obtained for expanded queries using pseudo relevance feed back method implemented in Terrier with default settings (T=10, D=3). Improvements relative to baseline are reported in the second part of Table 1. These results show that expanding queries by learning process is doing better than baseline however it doesn't reach PRF performance.

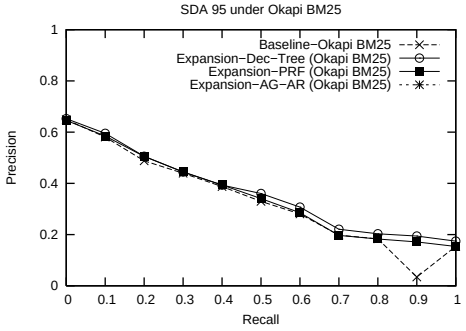

Figure 1: Recall/Precision curves of the expansion approaches for SDA95 collection.

Figure 1 shows evolution of the precision (P@11) for the different query expansion strategies.

6 CONCLUSION

We have presented in this article, a query expansion approach which is based on learning how to expand queries using association rules between terms. The problem of expansion is modeled as a supervised classification problem. An exploratory process based on genetic algorithms is used to both explore association rules space in search of the best terms for expansion and to generate training instances that are used later to build a classifier. The resolution of the learning problem is by decision tree. Experiments conducted on the French text collection SDA-95 show that learning how to expand queries using association rules is a promising approach. These preliminary results are encouraging, we plan to extend our representation of the input space by including new features in order to improve the learning process.

REFERENCES

Agrawal, R. and Skirant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the* 20^{th} *International Conference on Very Large Databases, VLDB 1994*, pages 478–499, Santiago, Chile.

Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., and Lakhal, L. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. In *Proceedings of the* 1^{st} *International Conference on Computational Logic*, volume 1861 of *LNAI*, pages 972–986, London, UK. Springer-Verlag.

Boughanem, M. and Tamine, L. (2000). Query optimization using an improved genetic algorithm. In *Proceedings of the 2000 ACM CIKM International Conference on Information and Knowledge Management, McLean, VA, USA, November 6-11, 2000*, pages 368–373.

Buckley, C., Salton, G., Allan, J., and Singhal, A. (1994). Automatic Query Expansion Using SMART: TREC-3. In *Proceedings of the* 3^{rd} *Text REtrieval Conference*.

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

Chifu, A. and Mothe, J. (2014). Expansion sélective de requêtes par apprentissage. In Moens, M., Viard-Gaudin, C., Zargayouna, H., and Terrades, O. R., editors, *CORIA 2014 - Conférence en Recherche d'Infomations et Applications- 11th French Information Retrieval Conference. CIFED 2014 Colloque International Francophone sur l'Ecrit et le Document, Nancy, France, March 19-23, 2014.*, pages 257–272. ARIA-GRCE.

Fonseca, B. M., Golgher, P. B., Pôssas, B., Ribeiro-Neto, B. A., and Ziviani, N. (2005). Concept-based interactive query expansion. In *Proceedings of the* 14^{th} *International Conference on Information and Knowledge Management, CIKM 2005*, pages 696–703, Bremen, Germany. ACM Press.

Ganter, B. and Wille, R. (1999). *Formal Concept Analysis*. Springer-Verlag.

Haddad, H., Chevallet, J. P., and 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 the* 4^{th} *European Conference on Principles and Practices of Knowledge Discovery in Databases, PKDD 2000*, Lyon, France.

Jones, K. S., Walker, S., and Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments. *Information Processing and Management*, 36(6):779–840.

Latiri, C., Haddad, H., and 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.

Lin, H. C., Wang, L. H., and Chen, S. M. (2008). Query expansion for document retrieval by mining additional query terms. *Information and Management Sciences*, 19(1):17–30.

Mitra, M., Singhal, A., and Buckley, C. (1998). Improving automatic query expansion. In *Proceedings of the* 21^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'98*, pages 206–214, Melbourne, Australia. ACM Press.

Pasquier, N., Bastide, Y., Taouil, R., Stumme, G., and Lakhal, L. (2005). Generating a condensed representation for association rules. *Journal of Intelligent Information Systems*, 24(1):25–60.

Pragati Bhatnagar, N. P. (2015). Genetic algorithm-based query expansion for improved information retrieval. In *Intelligent Computing, Communication and Devices, Advances in Intelligent Systems and Computing*, volume 308, pages 47–55.

Rungsawang, A., Tangpong, A., Laohawee, P., and Khampachua, T. (1999). Novel query expansion technique using apriori algorithm. In *Proceedings of the* 8^{th} *Text REtrieval Conference, TREC 8*, pages 453–456, Gaithersburg, Maryland.

Ruthven, I. and Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145.

Tangpong, A. and Rungsawang, A. (2000). Applying association rules discovery in query expansion process. In *Proceedings of the* 4^{th} *World Multi-Conference on Systemics, Cybernetics and Informatics, SCI 2000*, Orlando, Florida, USA.

Xu, J. and Croft, W. B. (1996). Query expansion using local and global document analysis. In *Proceedings of the* 19^{th} *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1996*, pages 4–11, Zurich, Switzerland. ACM Press.