2000

Mining Term Association Rules for Global Query Expansion: A Case Study with Topic 202 from TREC4

Jie Wie
National University of Singapore, weijie@comp.nus.edu.sg

Zhenxing Qin
National University of Singapore, qinzhenx@comp.nus.edu.sg

Stephane Bressan
National University of Singapore, steph@comp.nus.edu.sg

Beng Chin Ooi
National University of Singapore, ooibc@comp.nus.edu.sg

Mining Term Association Rules for Global Query Expansion: A Case Study with Topic 202 from TREC4

Jie Wei, Zhenxing Qin, Stéphane Bressan, Beng Chin Ooi
School of Computing
National University of Singapore
{weijie,qinzhenx,steph,ooibc}@comp.nus.edu.sg

Abstract

The sudden growth of the World Wide Web and its unprecedented popularity as a de facto global digital library exemplified both the strengths and weaknesses of the Information Retrieval techniques used by popular search engines. Most queries are short and incomplete attempts to describe or characterize the possible documents relevant to the query. It seems then natural to try and expand the queries with additional terms, which are semantically and/or statistically associated with the original query terms. In this paper we are looking at the mining of associations between terms for the exploration of the terminology of a corpus as well as for the automatic expansion of queries. The technique we use for the discovery of the associations is association rules mining [Agrawal 96]. The technique we propose is more flexible than previous techniques based on term co-occurrence since it takes into account not only the co-occurrence frequency but also the confidence and direction of the association rules. Our preliminary experiment results show we can get benefit from this novel technique.

1. Introduction

The growing volume of textual information readily available in digital libraries and on the Internet compels new techniques for the exploration of these large corpora and for the effective and efficient retrieval of documents relevant to the users' needs. In this paper, we discuss a novel yet straightforward approach to the mining of term associations and its application to the exploration of the corpus vocabulary and the expansion of queries. Our objective is therefore twofold.

First, we extract term associations, which, we claim, capture a contextual semantics stemming from the vocabulary usage in the corpus. Such associations are then gathered in a graph that the user seeking to understand the corpus vocabulary usage may explore interactively.

Subsequently, recognizing that most queries expressed by users are short and incomplete attempts to describe or characterize the possible relevant documents, we proceed to evaluate the effectiveness of query expansion based on the mined term associations.

The technique we use for the discovery of the associations is association rules mining [Agrawal 96]. It is more flexible than previous techniques based on term co-occurrence since it takes into account not only the co-occurrence frequency of terms but also the confidence and direction of the association rules.

For the discussion of our results we use the topic 202 of the TREC4 benchmark. Topic 202 applies to the collection of news wires issued by the Associated Press in 1990 (AP90).

For this research we have opted for the vector space model. A document is expressed as a vector whose coefficients are calculated from the occurrence and the frequency of the different terms inside one document (called the term-frequency). A query is a pseudo document. Among the many metrics available for formally defining the notion of similarity, we use a variant of the cosine similarity. We adopted the commonly used term weighting function $\ln(1/s(t))$. $s(t)$ is called the support of the term, the ratio of the total number of documents by the document frequency of a term t is called the support of the term. Several term weighting functions were discussed and studied in [Greif 1998].

In Section 2 we shortly discuss some query expansion techniques. In section 3, we describe our methodology and outline the architecture of our test-bed implementation. In section 4, we test the expansion with WordNet, a global thesaurus. In section 5, we informally argue that the graph of term associations conveys a contextual semantics and discuss the improvement and deterioration obtained with query expansion for various combinations of the parameters. To let you have an concrete idea of our expansion techniques, we show them with topic 202 from TREC4.

Finally we summarize our findings for topic 202 as well as the average results obtained over all the topics available for TREC4 and we indicate our current and future research directions.

2. Query Expansion and Thesauri

Whether they are well-formed affirmations of questions, or ungrammatical lists of words, queries in the information retrieval models typically are short and

incomplete sets of terms. Query expansion hypothesizes that the addition of terms to the query has the potential to overcome these weaknesses and improve the effectiveness of the retrieval. The empirical results from various authors show unfortunately that improvement is difficult to achieve and improvements of 5 to 10% are considered very significant [Voorhees 1993,1994].

Let us classify the expansion methods according to two criteria into four categories. An expansion method may be using local or global knowledge, i.e. look at the answers to a query or at the entire corpus. An expansion method may be using human input or be fully automatic.

Global methods use the entire corpus or external knowledge to expand the query. The knowledge used comes in the form of a thesaurus. A thesaurus is a data structure recording associations between terms.

Man-made Thesauri such as WordNet, whether they are general purpose of domain specific, records linguistic or conceptual associations between the terms as identified by human experts. In [Voorhees 1993,1994] the authors report various experiments with the WordNet thesaurus. In [Voorhees 1994] they manually expand the queries using all relations in the thesaurus: hyper/hyponyms and synonyms. After tuning of parameters and weights, they report up to 1.7 % improvement of the average precision (over several levels of recall). The experiments use a TREC3 corpus. In [Voorhees 1993] the authors only use the synonym relations. The expansion is automatic but applies to both the query and the documents.

Several authors have attempted to automatically mine the associations for query expansion from the corpus. Most approaches are based on clustering of terms in the document space. Intuitively, clustering captures synonymous associations. In [Lin 1998] the authors present a global analysis approach based on association rules mining. However, their objective is not to find associations for query expansion but rather to construct a classification of the documents.

The local set of documents is the set retrieved with an initial unexpanded query. Local methods use the local set to discover the additional candidate terms to be added to the query. In practice local set is kept arbitrarily small compared to the total size of the corpus. The user can indicate manually which documents in the local set are relevant. This step is called relevance feedback. The frequent words in the relevant documents can be used to expand the initial query and/or re-weight the query terms. The first positive result with such a technique was obtained by Salton as early as 1971 in the SMART system. Alternatively, associations of terms within the local set can be mined and chosen automatically by methods similar to the one we are studying in this paper. Similarly to global methods, most of the proposed techniques are based on clustering. In [Sanderson&Croft 1999] the authors use the confidence of the association between terms (they call the conditional probability) to construct a term hierarchy from the local set. They do not take the support of the terms and of the association into account. Although they indicate that such a hierarchy can be used for manual or automatic query expansion they do not further study the issue.

For an exhaustive overview of earlier work on local and global query expansion as well as thesauri construction see [Greif 1998].

The method we are studying is a global automatic query expansion method. There are few available studies on such methods for significantly large corpus, as the experiments are lengthy and greedy in memory utilization.

3. Methodology and system architecture

3.1 The Corpus and the Topics

We used AP90 from TREC4 as our benchmark. The AP90 corpus contains more than 78,000 documents. After stemming and the filtering of stop words, we found more than 133,000 different terms. The average number of different terms in a document is 217. The largest documents contain approximately 900 terms and the smallest 5 terms.

There are 50 queries or topics, called topics in the TREC terminology, associated with AP90. Each topic consists of a natural language proposition or question. For each of these topics, a list of relevant documents has been constructed manually and endorsed by human experts. This list is provided together with the topic. In average, a topic for AP90 calls for 32 documents. The largest topic calls for 145 documents. Two topics call no relevant document at all. Topic 202 is “Status of nuclear proliferation treaties -- violations and monitoring”. It calls for 52 documents.

3.2 Architecture and System Implementation

We use the Oracle 8 database system as a repository for the documents and indices we build. The application programs for the processing of documents and queries are written in Java. SQL statements are executed through JDBC to query the database.

As illustrated on Figure 1, the system is composed of four main modules: the indexing module, the association rule miner, the query expander, and the query processor. The common process of information retrieval is illustrated in the real line. The processes in dotted line are additional ones for the query expansion and rules presentation.

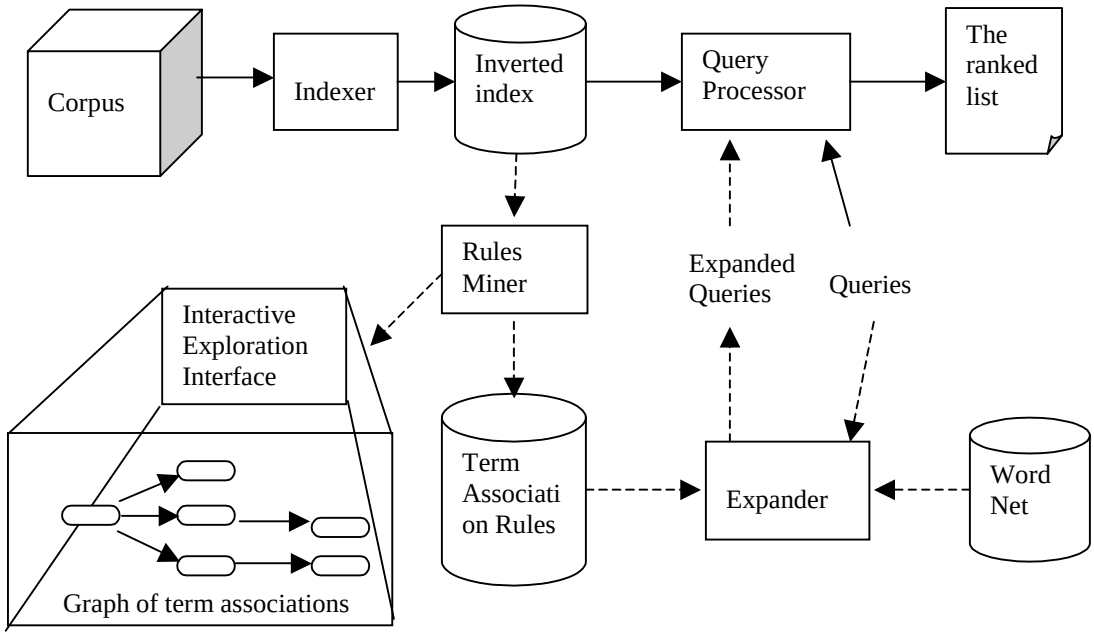

Figure 1

The indexing module constructs an inverted index of the corpus after tokenizing and stemming of the words in each document

The association rule miner extracts the term associations. The graph of terms thus constructed can be interactively browsed using the interactive exploration interface. The user browsing the graph of term associations can control the confidence and support parameters of the rules as well as the frequency of the terms she whishes to be displayed.

The query processor computes a weighted-cosine similarity between the query and the documents.

The query expansion module uses alternatively one of the term relation either constructed by the association rule miner or from WordNet (we translated the main WordNet relations into terms and stored them into Oracle.)

4. Thesaurus Expansion —WordNet.

4.1 WordNet

The WordNet thesaurus is an electronic lexical reference system for the English language whose design is inspired by current psycholinguistic theories of human lexical memory [Miller 1993,WordNet]. It has already inspired compatible developments in other languages [Euro WordNet], paving the road to multi- and cross-lingual applications.

In WordNet, English nouns, verbs, adjectives, and adverbs are organized into synonym sets called synsets. Each synset consists of a list of synonymous word forms, representing one underlying lexical concept. Semantic pointers describe relationships between the synsets. WordNet contains approximately 122,000 terms grouped into approximately 99,000 synsets and it is constantly being updated and extended. The synset represents the synonym equivalence relation. The main relations between synsets are the hyponym and hypernym , and meronym and holonym reciprocal relations.

peace

As an example we now give the hypernyms, hyponyms, and synomyms in the WordNet [Miller 1993,WordNet] thesaurus for the word “treaty”. The term “treaty” is a hypernym of the terms “alliance”, “peace treaty”, “peace”, “convention”, “SALTI”, and SALT II”. A convention is some kind of treaty. The term “treaty” is a synonym of the respective terms “pact” and “accord”. A treaty is an accord. It is an hyponym of “written agreement”. A treaty is some kind of written agreement.

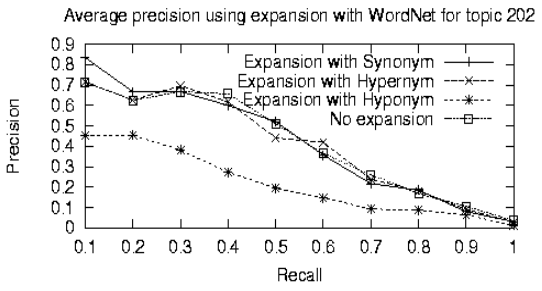

Figure 2

If we use the WordNet associations to expand the query, we do not observe consistent improvement. Figure 2 shows the effectiveness of the expanded query in comparison to the initial query.

The higher the corresponding plot the more effective the retrieval. The random performance has numerous causes, one important one being the fact that the general thesaurus, although semantically ideal, may not reflect the vocabulary, jargon, and usage within the corpus.

The cost of building a domain-and usage-aware thesaurus by hand and brain is probably so expensive that is not cost effective in most applications. Therefore we turn towards the automatic extraction of term associations.

5. Association Rules Expansion

5.1 Term association rules

A term association rule is a rule of the form $t_1 \Rightarrow t_2$ where t_1 and t_2 are terms. A rule is the product of the statistical analysis of a set of documents. It is characterized by both the confidence and the support.

The support of a term or a set of terms is the ratio of the number of document in which the term or each term of the set appears to the total number of documents.

The confidence of a rule $t_1 \Rightarrow t_2$ is the ratio of the support of the rule to the support of t_1:

$c = s(\{t_1, t_2\}) / s(t_1)$

A rule with a high confidence indicates that term t_2 often occurs in a document where term t_1 occurs. A rule with high support indicates that many examples were found to suggest the rule.

The scope of our investigation is defined by the variation of the range of the two parameters and by the options in using rules of one or more of the forms below given for the query term "nuclear":

(1) "nuclear" => t s, c ;
(2) t => "nuclear" s, c;
(3) t <=> "nuclear", i.e. t => "nuclear" and "nuclear" => t, s, c1,c2

Of course such rules exist as soon as a term t appears in a document where the term "nuclear" appears. We use the confidence and the support of the rules, indicator of their quality, to select some of them only.

For example and for some ranges of support and high confidence, rules of the type (1), (2), and (3), suggest the terms, respectively:

```
nuclear=>soviet,
  (supp=0.016,conf=0.4872)
nuclear=>U.S.,
  (supp=0.0187,conf=0.5688)
plutonium=>nuclear,
  (supp=0.0017,conf=0.8993)
reactor=>nuclear,
  (supp=0.0039,conf=0.9711)
weapon<=>nuclear,
  (supp=0.0171,conf1=0.2825,
                    conf2=0.5194)
```

5.2. Is there a Natural Semantics behind the Rules?

A high confidence rule of the form $t_1 \Rightarrow t_2$ indicates that (often) t_2 appears in a document if t_1 appears. This suggests hyper/hyponym or holo/meronym types of relations between the terms. t_2 is equally or more general than t_1 and conversely. Examples mined from our corpus are "Kohl" is a "Chancelor", and "soybean", "corn", and "wheat" are kinds of "grain". The relations found characterize what we could call a contextual holonony: if $t_1 \Rightarrow t_2$, then t_1 is part of the vocabulary in the topical context suggested by the concept denoted by t_2. For example we found such associations between the names of "Mandela" and "De Klerk" with the term "Apartheid".

A rule t1 $\Leftrightarrow$ t2, i.e. t1 => t2 and t2 => t1 with high and similar respective confidence as well as a high support indicates that t1 and t2 tend to appear together. This suggests a contextual synonymy. An example is the mined association between "conviction" and "sentence".

Many associations were also mined between nouns and adjectives not handled by the stemming algorithm such as "Jew" and "Jewish", "Japan" and "Japanese". There are also many such associations associating the first and last names of personalities: "Saddam" and "Hussein", "Mikail" and "Gorbachev" (provided the first and last name or not ambiguous in the corpus). Of course the synonymy is not proper. The association indicates the similarity of the terms in the topical context such as the association found between the two terms "crisis" and "Gulf" in the 1990 corpus.

Although we have not conducted a formal evaluation of the results obtained, we notice that our subjective evaluation leads to similar conclusion than those presented and substantiated in [Sanderson 1999]: it is possible to obtain a topically meaningful thesaurus from the statistical analysis of a corpus. In our case the meaning of the association discovered can be related to the form and parameters of the association rules. A formal evaluation of the semantic quality of the rules could be conducted for instance with a protocol to ask users

familiarized with the corpus or the corpus subject matter to assess the rules.

5.3. The performance of query expansion with association rules

Figure 3 to 5 show the performances of the corresponding expansion in comparison to the initial query for our example. Despite the striking subjective semantic relevance of the associations mined, the improvement or the deterioration we observe are not consistent over the entire set of queries we have been testing. The results are also very sensitive to the variation of the ranges and threshold chosen for the confidence and support parameters.

5.3.1 Expansion with rule Q=>X. (Confidence>0.4, 0.001<Support<0.001).

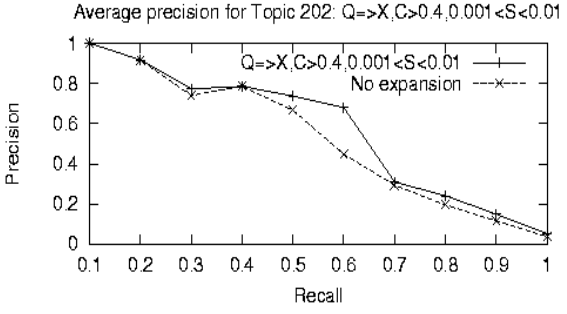

Figure 3

Figure 3 shows the performance of expansion with terms implied by the query terms for topic 202. The confidence and support parameters are chosen experimentally, but applied to all the other topics. We set a low-bound of support 0.001. Rules with support below this bound are thought appear by casual. We select the support bigger than 0.001 i.e. the query words and expanded words co-occurred in at least 0.1% of the documents. Since the AP90 is topic diverse, this support is not low.

The expanded words are "**united, states, US, year, weapon**". Many of them are frequently appeared words. These frequently appeared words have low weight in the document vector, so they affect the cosine value not much. Only "weapon" is not a frequent word here. So this is like the performance of method in 5.3.3.

5.3.2. Expansion with Rule X=>Q. (Confidence>0.8, Support>0.001)

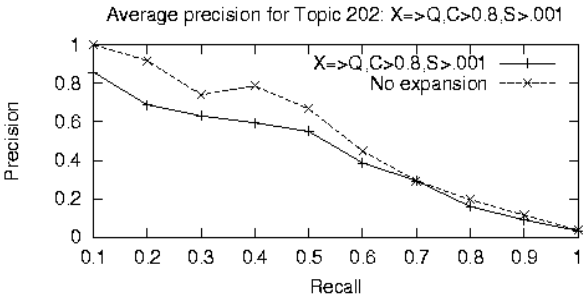

Figure 4

Figure 4 shows the performance of expansion with words implying query terms. The query words are relatively common words, i.e. have high support. In a rule, it's very often that low support word implies high support word when choose a high confidence threshold (we choose 0.2). In this method, the candidates of expansion words have low support and there are too many such words. We set a higher confidence here than in the former one is to limit the number of words expanded.

The expanded words are "**plutonium, reactor, warhead**". This method degraded the performance consistently. Words expanded through this way are mostly specific to the query words. In our intuition, specific words should perform better than more general meaning words. On the other hand, expansion with specific words is more risky than general meaning words. This is because many words expanded here are specific to the query words but not to the query itself.

5.3.3. Expansion with Rule X⇔Q. (Confidence1, Confidence2>0.25, Support>0.001)

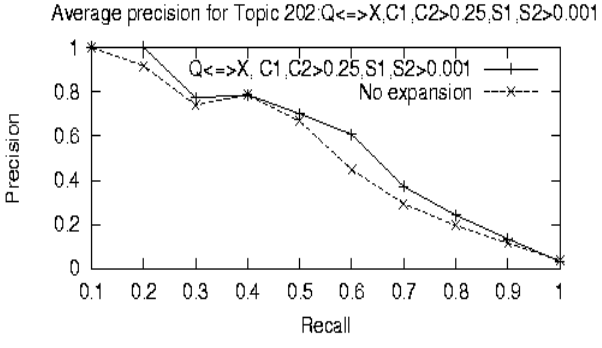

Figure 5

Here we expand words that implying and implied by query words with both confidence bigger than 0.25. We think these words expanded with both rules direction are context synonyms to query words. In such rules, the terms in the two sides have the same scale of supports. In this topic, only word "**weapon**" is in expansion set. Here "weapon" is implying and implied by query term "nuclear". It is a very meaningful word for topic 202 and appears in all the relevant documents to such topic.

6. Conclusion and Future Work

Using the mined association rules, we have been able to consistently improve the effectiveness of the retrieval over the set of 48 topics available for the AP90 (Figure 6). Using rules of the form Q=>X with lower and upper bounds on the support for rules we obtained a consistent improvement at all recall points. [J.Wei 2000]

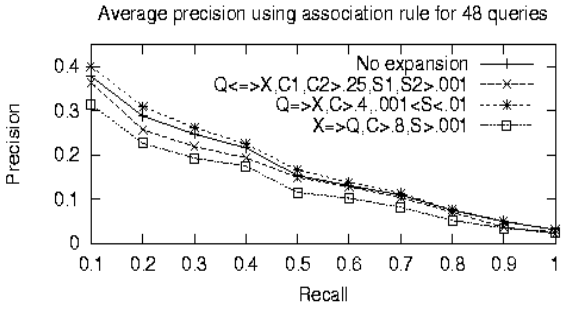

Figure 6

However, our results are preliminary. The tuning of the parameters for the selection of the rules, the ranges of the support and the confidence, was done in a non-systematic way due to the duration of each experiment. Such an approach is however made possible for association rules thanks to the additional control parameters (confidence and direction) not available in the previously proposed methods based on the sole term co-occurrence. Encouraged by a rare performance increase of up to 5%, we are now embarking on the systematic evaluation of the influence of the parameters.

The next objective of this research is to mine term associations for corpus written in languages for which Thesaurus such as WordNet are not readily available or too expensive. We wish to study text information retrieval and query expansion for languages such as Mandarin and Malay/Indonesian and with, respectively, two hundred and twenty million and one billion native speakers joining the information society.

References

A.F. Smeaton and I.Quigley, "Experiments on Using Semantic Distances Between Words in Image Caption Retrieval.", *International Conference on Research and Development in Information Retrieval*, 1996.

G.A.Miller, R.Beckwith, C.Felbaum, D.Gross and K.Miller, "Introduction to WordNet: An On-line Lexical Database.", Revised Version 1993.

Ellen M. Voorhees and Yuan-Wang Hou, "Vector Expansion in a Large Collection", *First Text REtrieval Conference (TREC-1)*, 1993.

R.Richardson, A.F.Smeaton and J.Murphy, "Using WordNet as a Knowledge Base for Measuring Semantic Similarity between Words."

E.M.Voorhees, "Query Expansion using Lexical-Semantic Relations", *ACM-SIGIR*, 1994.

W.R.Greiff, "A Theory of Term Weighting Based on Exploratory Data Analysis", *ACM SIGIR*, 1998.

S.H. Lin, C.S. Shih, M.C. Chen, J.M. Ho, M.T. Ko and Y.M. Huang, "Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A Semantic Approach.", *ACM SIGIR*, 1998

M.Sanderson and B.Croft, "Deriving Concept Hierarchies from Text.", *ACM SIGIR*, 1999

Agrawal, R., Imielinski, T., and Swami, A., "Mining Association Rules between Sets of Items in Large Databases.", *Procedings of the ACM SIGMOD*, 1996

EuroWordnet, "Building a Multilingual Database with Wordnets for several European Languages", www.hum.uva.nl/~ewn/

Wordnet, "WordNet - a Lexical Database for English.", www.cogsci.princeton.edu/~wn/

W. Greif, "A Theory of Term Weighting Based on Exploratory Data Analysis.", *ACM SIGIR*, 1998

R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval.", 1999

J. Wei, S.Bressan, B.C.Ooi, "Mining Term Association Rules gor Automatic Global Query Expansion: Methodology and Preliminary Results", *the first International Conference on Web Information Systems Engineering*, WISE 2000