Association-Based Segmentation for Chinese-Crossed Query Expansion

Chengqi Zhang, Zhenxing Qin, Xiaowei Yan

Abstract—**The continually and high-rate growth of China's economy has attracted more and more international investors. These investors have an urgent need of identifying patterns in Chinese information, which are potentially useful in making competitive decisions. The first step of deeply understanding and analyzing Chinese information is how to effectively search those likely relevant to a user query. However, queries provided by users are often incomplete and inappropriate to the information systems, especially for retrieving Chinese-crossed information. In this paper, we present a segmentation based on actionable Chinese term-association analysis for better understanding user queries so as to automatically generate Chinese-crossed-query expansions. The semantics behind the actionable term-association rules is thus studied. Experiments conducted have shown that our approach is efficient and promising.**

Index Terms—**Chinese and Japanese characters, Information retrieval, Text processing,**

I. Introduction

The pressure of enhancing corporate profitability has caused companies to spend more time on identifying diverse sales and investment opportunities for winning China's markets. The short list of examples below should be enough to place the current situation into perspective (these examples are cited from [12]):

- The year 2008 may seem like a long way away, but commercial enterprises know that now is the time to line up for sponsorship of Beijing's Olympic Games. It's the moment China has been dreaming about for years. It is a time of national pride and celebration and also a time of opportunity. Many international companies see the 2008 Olympics as a chance to earn huge profits. (see http://www.creadersnet.com/newsViewer.php?idx=99079)
- Credit Suisse-First Boston (CSFB) says it has listed the China market as one of its primary targets for business development and is focusing its global resources on the country. John J. Mack, CEO of the CSFB, told that China's continued economic reforms and its entry into the World Trade Organization have fostered great potential for business development. (see http://www.creadersnet.com/newsViewer.php?idx=129913)
- PeopleSoft, the fifth largest software maker in the world, will invest in China in the coming eight years. "We make a few investments every year and China will be our biggest move this year in terms of market and products," Craig Conway, chief executive officer and president of PeopleSoft. He said China's accession to the World Trade Organization (WTO) means more local enterprises will have chances to compete with their international counterparts and are anxious to upgrade their management with enterprise software, so PeopleSoft is coming at a "prefect time". (see http://www.creadersnet.com/newsViewer.php?idx=129912)

With the increasing interest to China's markets, the need to be able to digest the large volumes of Chinese information is now critical. In particular, it is very important to discover and develop Chinese ancient cultures and medicines, which brings benefit to mankind. The first step of deeply understanding and analyzing these Chinese information is how to effectively search those likely relevant to a user query. Accordingly, this paper focuses on the issues of Chinese information retrieval.

User queries to the Web or other information systems are commonly described by using one or more terms as keywords to retrieve information. Some queries might be appropriately given by experienced and knowledgeable users, while others might not be good enough to ensure that those returned results are what the users want. Some users consider that Boolean logic statements are too complicated to be used. Usually, users are not experts in the area in which the information is searched. Therefore, they might lack the domain-specific vocabulary and the author's preferences of terms used to build the information system. They consequently start searching with generic words to describe the information to be searched for. Sometimes, users are even unsure of what they exactly need in the retrieval. All of these reasons then often lead to uses of incomplete and inaccurate terms for searching. Thus, an information retrieval system should provide tools to automatically help users to develop their search descriptions that match both the need of the user and the writing style of the authors.

One of the solutions to provide the service is the automatic expansion of the queries with some additional terms [1, 2]. These expanded terms for a given query should be semantically or statistically associated with terms in the original query.

Manuscript received May 9, 2004. This work is partially supported by a large grant from Australian ARC (DP0559536), a China NSFC major research Program (60496321), and a China NSFC grant (60463003)

Chengqi Zhang is with the Faculty of Information Technology, University of Technology Sydney, PO Box 123, Broadway, Sydney NSW 2007, Australia. (e-mail: Chengqi@it.uts.edu.au)

Zhenxing Qin is with the Faculty of Information Technology, University of Technology Sydney, PO Box 123, Broadway, Sydney NSW 2007, Australia. (e-mail: zqin@it.uts.edu.au).

Xiaowei Yan is with Department of CS, Guangxi Normal University, Gulin, China. (e-mail: yanxw@mailbox.gxnu.edu.cn)

Moreover, techniques of association rule mining [9, 10] are frequently used for text mining [3, 5, 6, 8] and global query expansion [4, 7].

For Chinese information, users often search for information with Chinese-crossed queries. Because of the complicated morphology, syntax and semantics of Chinese language, it is very difficult to generate the efficient and intelligent service of Chinese-crossed text retrieval.

In this paper, we introduce our system for dealing with Chinese-crossed queries, in which a segmentation based on actionable Chinese Term-Association Rule (CTAR) analysis is proposed for better understanding user queries so as to automatically generate Chinese-crossed-query expansions. In Section 2, the system structure, the actionable CTAR analysis and the establishment of thesauri are depicted. In Section 3, performance of our method is evaluated based on our experiments.

II. Problem Statement

A. System Structure

Figure 1 illustrates the system structure that we design for Chinese-crossed query expansion.

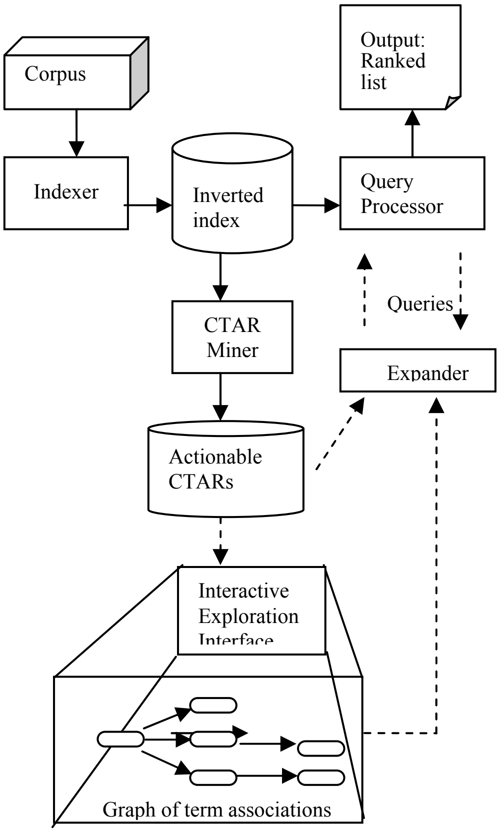

Figure 1: System structure for Chinese-crossed query expansion.

The system consists of four main parts: indexing module, actionable CTAR miner, query expander and query processor. The indexing module constructs an inverted index for a corpus. Unlike English text processing, Chinese text processing does not need tokenizing and stemming, but a segmentation step is required to parse the Chinese text for a term list. We will discuss the segmentation later.

The actionable CTAR miner extracts those Chinese term-association rules actionable to Chinese-crossed query expansion. The graph of term-associations, thus constructed, can be interactively browsed by using the interactive exploration interface. User can control the confidence and support thresholds of actionable CTARs, as well as the frequency of the terms to be displayed, by browsing the graph of term-associations.

The query processor calculates a weighted-cosine similarity between a query and a document.

The query expansion module uses the term-association and Chinese Thesauri to expand queries.

B. The Segmentation Process

Chinese text is different from English text. Namely, there is no explicit word boundary in a sentence of Chinese. In English text, words are separated by spaces or punctuations. We can easily extract the terms from English text. In Chinese text, words are made up of one, two or more Chinese characters, and the same character can occur in many different words. The separators, such as comma and period, are only found between sentences. Furthermore, there is no explicit indication to tell where one word begins or ends in a Chinese sentence.

Decomposing a Chinese sentence into many single Chinese characters is not a good approach of segmentation. In Chinese, some words are composed of the same characters, but have different meanings in different sequences. Query expansion is a useful approach to improve the recall of retrieval. No matter whether you are expanding queries with a thesaurus or doing global/local analysis, you should work on substantive words. To apply the expansion methods to Chinese text retrieval, we need extract the words from sentences. How to exact words from Chinese text is always a challenge task in Chinese information retrieval. The process of breaking sentences into words is called segmentation.

There are two major segmentation techniques. One is the statistical approach and another is the dictionary-based approaches. The first method works well in finding bigrams. But its limitation is that it can only deal with words not longer than 2 characters. In our expansion system, the second segmentation method is chosen. It uses a lexicon tool (a Chinese wordlist) to find the word boundaries. It is more flexible and proper for query expansion.

The segmentation method exactly used in our system is called the forward maximum matching method. The segmentation is a Marlkov process, because the next word from segmentation is only decided by the current sentence and the dictionary, being unrelated with the words segmented before.

We first get a set of sentences from a Chinese text using the nature segmentation symbols, such as spaces, punctuations. For each sentence, we use forward maximum matching method with a dictionary finding the boundaries between words.

The dictionary we used contains 44,000 words with the length varying from 1 to 7 characters. We first put these words into 7 catalogues according to their length and sorted them meanwhile. The sorting is to facilitate the later searching. They can be read into memory during the segmentation process. These 7 catalogs are named wordlist1 to wordlist7 according to the length of the words inside the catalogues. The segment algorithm is shown below.

```
Segment (sentence)
Input: sentence, wordlist1, … , wordlist7
Output:  a list of words

Step 0.  word_list ← empty ;
Step 1.  $l$ ← length of the sentence
Step 2.  if  $l = 0$ then stop and output word_list;
Step 3.  if $l > 7$,  $l \leftarrow 7$;
         prefix ← sentence (0, $l$)

Step 4.  if prefix can be found in wordlist$l$ then
       { add prefix to word_list;
         sentence ← sentence - prefix;
         goto step1}
     else if  $l>2$  { $l \leftarrow l - 1$; goto Step3}
         else { prefix ← sentence (0,1);
              add prefix to the output wordlist;
              sentence ← sentence-prefix;
              goto step1; }
```

C. Actionable Chinese Term-Association Rules

We set the minimum support at 0.0001 and minimum confidence at 0.1 in the rules mining. These thresholds are smaller than those for English corpus, because the Chinese words are distributed more evenly in different documents. Figure 2 shows the distribution of words frequency for a Chinese corpus introduced later. The horizontal axis represents document frequency in e~k. The vertical axis is the ratio of words, which fall into the document frequency. From the most-left point in Figure 2, we can see that about 10% of words occur in less than e1 documents. In AP90 from TREC4, the collection of news wires issued by the Associated Press in 1990 [7], there are 45% of words fall in this field. In this English corpus, we eliminate the stop words in indexing process, and there are still 455 words appearing in at least 5000 documents. In the Chinese corpus, on the other hand, we do not pick out stop words, and there are only 300 words appearing in more than 5000 documents. Because of this distribution over word frequency, there are less words co-appearing in the same documents. Therefore, we choose a smaller confidence threshold in Chinese CTARs mining, and later in rules selection in query expansion.

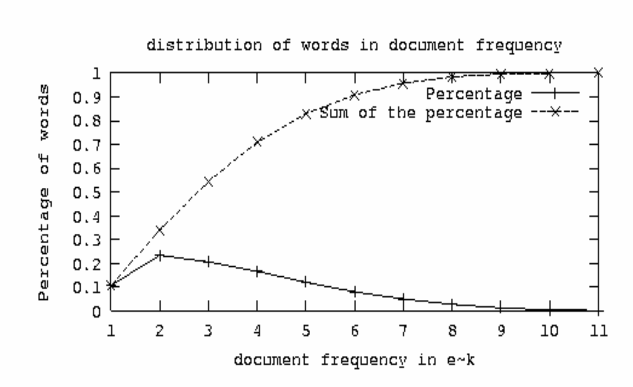

Figure 2: distribution of words frequency for a Chinese corpus.

D. Natural Semantics behind Actionable CTARs

A high confidence rule of the form t1 => t2 indicates that t2 often appears in a document if t1 appears. This suggests a certain type of relations between the terms, such as hypernym/hyponym or holonym/meronym, indicating a narrower/broader meaning between the terms. These relations characterize what we could call a contextual holonony, that is, if t1 => t2, then t1 is part of the vocabulary in the topical context, which is suggested by the concept denoted by t2. We categorized such rules into following four classes. The categorization is motivated from our inspection. Although this classification does not always seem reasonable, it can help us to understand the latent semantics behind those rules and their effects on the later query expansion.

(1) Hypernym/Hyponym ("a kind of" relation)

"Weltanschauung" is a kind of "ideology". "Balance beam" is a kind of "gymnastics". And "Marxism" is a kind of "theory". The corresponding rules are

世界观 (weltanschauung) => 思想 (ideology)

SUPP = .00089, CONF = .82258

平衡木 (balance beam) => 体操 (gymnastics)

SUPP = .00027, CONF = .8421

马克思主义 (Marxism) => 理论 (theory)

SUPP = .00125, CONF = .8372

(2) Meronym/Holonym ("a part of" relation)

"Bremen" is a part of Germany. "Ophthalmology" is a part of the hospital. "Wing" is a part of a aeroplane.

不莱梅 (Bremen) => 德国 (Germany)

SUPP = .00017, CONF = .90909

眼科 (ophthalmology) => 医院 (hospital)

SUPP = .00012, CONF = .875

机翼 (aerofoil) => 飞机 (aeroplane)

SUPP = .00034, CONF = 1

(3) Other narrower/broader meanings in topics

Sometimes, words in the rules are not strictly in hypernym/hyponym or meronym/holonym relationship. But they indicate a relationship of narrower/broader meaning in topic. "Sentence" is a narrower topic than that of "law". The other example is shown by cancer and hospital.

量刑 (sentence) => 法律 (law)

SUPP = .00015, CONF = .81818

羊城 (city of sheep) => 广州 (GuangZhou)

SUPP = .00017, CONF = .83333

侨胞 (emigrant) => 海外 (oversea)

SUPP = .00122, CONF = .85365

肿瘤 (cancer) => 医院 (hospital)

SUPP = .00013, CONF = .88888

(4) Special word usage in Chinese

In Chinese, there are some special word usages in verb-object group, subject-verb group and adjective-noun group. In verb-object group, for example, when mentioning a certain objects we mostly uses the corresponding verb, or (not and) vice versa. An example is that when an ambassador handover credentials, the verb for handover in Chinese must be "递交".

For subject-verb groups there is a similar situation. If "喝采" (applause) occurs, the subject is likely "观众" (audience).

Some adjectives are designated to modify specific nouns. For instance, "翻天覆地" (turn over the world) is specific for "变化" (change). "悠久" (age-old) is mostly used to modify "历史" (history). The related rules are as follow.

国书 (credential) => 递交 (handover)

SUPP = .00031, CONF = .81818

喝采 (applause) => 观众 (audience)

SUPP = .00022, CONF = .8125

决口 (breach) => 洪水 (flood)

SUPP = .00027, CONF = .8421

翻天覆地 (turn over the world) => 变化(change)

SUPP = .00076, CONF = .93617

悠久 (age-old) => 历史 (history)

SUPP = .00055, CONF = .94117

A rule t1 <=> t2, i.e. t1 => t2 and t2 => t1 with high and similar respective confidences, conf1 and conf2 respectively, as well as a sufficient support, indicates that t1 and t2 tend to appear together. We refer to t1 and t2 as “context synonyms”. Several kinds of “context synonyms” are given below.

(1) Synonym

t1 and t2 are real synonyms. For example, "防汛" (flood prevention) and "抗洪" (fight a flood), "漏税" (evade taxation) and "偷税" (evade taxes), "腐朽" (rotten) and "侵蚀" (erode), "检举" (report an offense to the authorities) and "揭发" (expose a crime) are all synonymic pairs. Here also list some of the rules.

防汛 (flood prevention) <=> 抗洪 (fight a flood)

SUPP = .0008, CONF1 = .32857, CONF2 = .34586

断流 (a river stops flowing) <=> 干涸 (dry up)

SUPP = .00012, CONF1 = .46666, CONF2 = .4375

漏税 (evade taxation) <=> 偷税 (evade taxes)

SUPP = .00012, CONF1 = .36842, CONF2 = .30434

腐朽 (rotten) <=> 侵蚀 (erode)

SUPP = .00061, CONF1 = .47945, CONF2 = .59322

检举 (report offenses to authorities) <=> 揭发 (expose crimes)
SUPP = .00026, CONF1 = .42857, CONF2 = .6

姑息 (tolerate evil) <=> 迁就 (yield to)

SUPP = .00026, CONF1 = .42857, CONF2 = .68181

受贿 (accept bribes) <=> 贪污 (corruption)

SUPP = .00075, CONF1 = .46236, CONF2 = .42574

(2) Antonym

It is an interesting relation. In Chinese, antonyms are often used together to emphasize something, such as extensive and intensive, modulation and demodulation.

粗放 (extensive) <=> 集约 (intensive) (in farming or management)

SUPP = .0008, CONF1 = .34328, CONF2 = .33823

调制 (modulation) <=> 解调 (demodulation)

SUPP = .00043, CONF1 = .64102, CONF2 = .92592

男生 (schoolboy) <=> 女生 (schoolgirl)

SUPP = .00012, CONF1 = .63636, CONF2 = .41176

(3) Peers relation.

Some closely-related peers appear together, such as "springboard" and "platform" in diving, "uneven bars" and "balance beam" in gymnastics, "Franc" and "Pound" in currency, "Germany" and "France" in countries, etc.

跳板(跳水) (springboard diving) <=> 跳台(跳水) (platform diving)

SUPP = .00047, CONF1 = .58695, CONF2 = .4909

高低杠 (uneven bars) <=> 平衡木 (balance beam)

SUPP = .0002, CONF1 = .46153, CONF2 = .63157

法郎 (franc) <=> 英镑 (pound)

SUPP = .00323, CONF1 = .42923, CONF2 = .73122

德国 (Germany) <=> 法国 (France)

SUPP = .01813, CONF1 = .33977, CONF2 = .39694

(4) Country and its capital

Though capital is a part of a country and the real relation between them should be viewed as meronyms/holonyms. They often appear together in context. When an author mentions a country, its capital often appears in the document and vice versa. An example is that between 贝鲁特 (Beirut) and 黎巴嫩 (Lebanon).

黎巴嫩 (Lebanon) <=> 贝鲁特 (Beirut)

SUPP = .00234, CONF1 = .5214, CONF2 = .64734

泰国 (Thailand) <=> 曼谷 (Bankok)

SUPP = .00146, CONF1 = .37004, CONF2 = .50299

(5) People and their workplace

Some people work in certain places. For example, 民警 (policeman) works in 派出所 (local police station), 渔民 (fisher) works in渔船 (fishing vessel), etc. Please look at the rules below for more details.

民警 (policeman) <=> 派出所 (local police station)

SUPP = .00083, CONF1 = .48, CONF2 = .41379

渔民 (fisher) <=> 渔船 (fishing vessel)

SUPP = .00031, CONF1 = .54545, CONF2 = .52941

守门员 (goalkeeper) <=> 球门 (goal)

SUPP = .00043, CONF1 = .38461, CONF2 = .36231

农民 (farmer) <=> 农村 (countryside)

SUPP = .01193, CONF1 = .42554, CONF2 = .48234

(6) Special word usage in Chinese

We already mentioned some special word usage in Chinese previously. Some words in the group depend on each other and are often used together. A subject-verb example is 候选人 (candidate) and 竞选 (election contest). A verb-object example is 体察 (observe) and 民情 (condition of the people). There are some phrases also can be considered in this category, such as " 不仅" (not only) and "而且" (but also), "通俗" (popular) and " 易懂" (easy to understand).

候选人 (candidate) <=> 竞选 (election contest)

SUPP = .00207, CONF1 = .40067, CONF2 = .40202

体察 (observe) <=> 民情 (condition of the people)

SUPP = .00015, CONF1 = .5625, CONF2 = .40909

通俗 (popular) <=> 易懂 (easy to understand)

SUPP = .00027, CONF1 = .55172, CONF2 = .8421

不仅 (not only) <=> 而且 (but also)

SUPP = .03548, CONF1 = .53098, CONF2 = .60464

(7) Local term-associations

There are also some relations only held in the corpus like those in English corpus. 海峡 (strait) and 台湾 (Taiwan) is an example. In Chinese news, 海峡 (strait) is often referred to Taiwan Strait and vice-versa.

海峡 (strait) <=> 台湾 (Taiwan)

SUPP = .00365, CONF1 = .50483, CONF2 = .34431

Examples of rules shown above have relatively high confidence, but not necessarily. Some rules with lower confidences (≈ 0.1) are still meaningful. For example,

结婚 (marry) <=> 婚礼 (wedding)

SUPP = .00015, CONF1 = .3, CONF2 = .10227

III. Performance of Chinese Query Expansion with Actionable CTARs

After exploring the semantic meaning of actionable CTARs in last section, we now tried to use these rules in query expansion.

A. Corpus and Queries

The corpus is a collection of news from the Xinhua News Agency in 1990. It has 57,240 documents. We segment the texts into Chinese words by using a dictionary. We do not eliminate stop words by using a stop-list. The documents are indexed after segmentation. There are 39,122 distinct words appearing in this corpus. The length of documents ranges from 5 to 3558 words, with 321 words on average.

We use 10 queries from TREC5 Chinese queries in our experiment, as listed below. E-title is the English translation from Chinese query. C-title is the original Chinese query, being used in retrieval.

1. <E-title> Communist China's position on reunification

<C-title> 中共对于中国统一的立场

2. <E-title> The newly discovered oil fields in China.

<C-title> 中国大陆新发现的油田

3. <E-title> Regulations and Enforcement of Intellectual

Property Rights in China

<C-title> 中国有关知识产权的立法与政策以及执法情况

4. <E-title> Numeric Indicators of Earthquake Severity in Japan

<C-title> 地震在日本造成的损害与伤亡数据

5. <E-title> Drug Problems in China

<C-title> 中国毒品问题

6. <E-title> World Conference on Women

<C-title> 世界妇女大会

7. <E-title> The Debate of UN Sanctions Against Iraq

<C-title> 联合国对伊拉克经济制裁的辩论

8. <E-title> The Mid-East Peace Talks

<C-title> 中东和平会议

9. <E-title> Measures to Prevent Forest Fires in China

<C-title> 中国森林火灾的防范措施

10.<E-title> Robotics Research in China

<C-title> 中国在机器人方面的研制

Because the corpus we use is not from TREC5, we can not obtain answers for the above queries from TREC5. It is impossible for us to read all the news and find out the exact precisions and recalls. Instead, we retrieve the top m documents for each query, and estimate the relevance of these documents. We then compare the average precision of these top m documents.

B. Expansion Methods and Retrieval Model Used

We tried three expansion methods based on three kinds of directions of rules, namely query words to expanded words, expanded words to query words, and dual directions.

We choose the parameters of support and confidence and varied them for each method, and select the best expanded terms (the expanded words most related to the queries). Then we retrieved the documents by four kinds of queries, the original query and the queries expanded with the three expansion methods.

The retrieval model is the Vector Space Model. We use the same similarity formula as that in [3]. The returned documents are sorted by their similarity values in descending order. We only consider the top m documents, ranked from 1 to m.

C. Relevance Estimation and Precision Value Computation

As discussed before, we retrieve documents by using a query and their three expanded queries respectively. The first m documents in the ranking list are selected for each query. Hence, we have at most 4*m documents for each query. Actually, there are many documents are retrieved, but only no more than 4*m documents are chosen for estimation for each query. In our experiment, when m = 10, there are 17.9 documents on average to be read for each query.

After retrieval, the retrieved documents are sent to a group of persons for estimation of the relevance to the query. The group of readers are all Chinese native speakers, but not experts on information retrieval and. We use the readers' votes as the value of precision. If there are n users who mark the first-ranked document relevant to a query, we say that precision of the document is n/10. The average precision of a query is the average precision of m documents for the query. The average precision of a rank is the average precision of the documents with the rank for all the 10 queries. The performance of average precision on each rank is shown below.

D. Comparison of Expansion Methods

We now compare the three expanded queries with the original query respectively. The query of No. 9, “中国森林火灾的防范措施” (Measures to Prevent Forest Fires in China), is used to illustrate our comparison.

(1) Expansion with Rules Q=>X

The expanded terms are relatively common words with high frequency. The average document frequency is about 4700. From the distribution of words in document frequency as shown in Figure 1, a word with this frequency (between e8 and e9) is among the top 5% frequent words.

Although we choose a small confidence of 0.2, it's still difficult to expand query terms which have high document frequency. This method favours those query terms with relatively lower document frequency, that is, those queries are easy to be expanded with general meaning words.

For the query of "Measures to Prevent Forest Fires in China", the expanded words are 生态(ecology), 保护 (protect), 环境 (environment), 林业 (forestry), 面积 (area), 检查 (inspect), 安全 (safety), 犯罪 (crime), 案件 (law case), etc. The corresponding rules are

森林 (forest) => 生态 (ecology), 保护 (protect), 环境 (environment), 林业 (forestry), 面积 (area)

火灾 (fire disaster) => 检查 (inspect), 安全 (safety), etc.

防范 (prevent) => 犯罪 (crime), 案件 (law case)

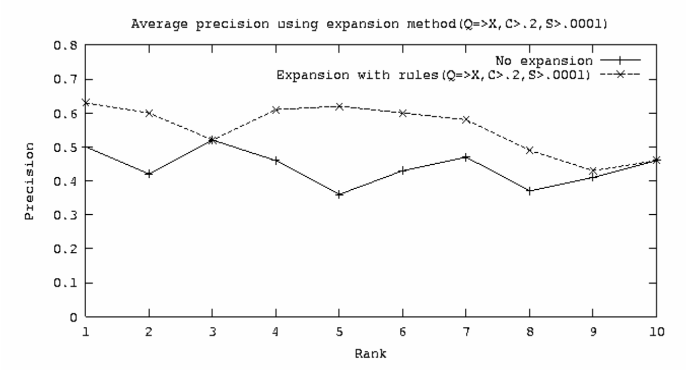

Figure 3: The average precision for each rank with rules Q=>X

The average precision over the ten ranks for this query is improved from 0.45 to 0.78 or 73.33%. Similarly, we expand all the 10 queries. The improvement in average precision is 25.91% (from 0.44 to 0.554). Performance of the query No. 2 is a little worse than that of original query. Performance of query No. 3 gains less than 5% of improvement. Query No. 5 has more than 10% improvement in performance. The expansions improve the precision in all ranks, except the third and tenth ranks where the performance keeps the same. Figure 3 shows the average precision for each rank.

(2) Expansion with Rules Q<=X

This method expands the query with words which imply occurrence of the query terms. On the contrary to the method of Q=>X, this method favours the relatively common query terms which have high document frequency, and expand the original query with specific meaning words.

We choose a higher confidence and support here to reduce the amount of expanded words and increase the quality of those words. For the query of No. 9, only one word is expanded, i.e. 防火 (fire prevention). The rule is

防火 (fire prevention) => 火灾 (fire disaster).

After expansion, the average precision is promoted from 0.45 to 0.67 or 48.89%. Because of the strict choose of parameters, only 8 queries out of 10 are expanded. When computing the overall average precision, we use the precision of original query for the two unexpanded queries. Improvement of the overall average precision is 27.95% (from 0.44 to 0.563). We only obtain significant improvement in the performances with two queries, but no decreased performance is found.

When look into the 10 ranks, there is only a little decrease at rank 6th. The other ranks all get benefit. See Figure 4.

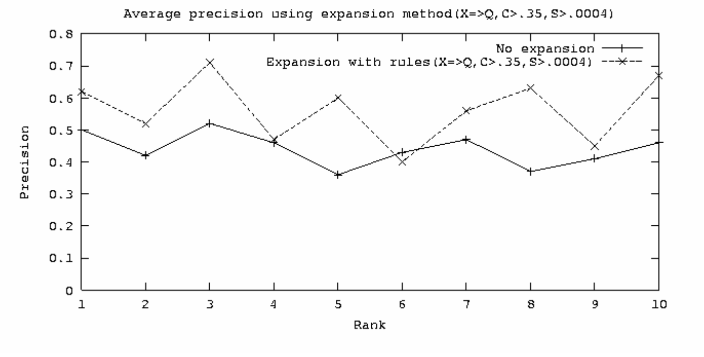

Figure 4: The average precision for each rank with rules Q<=X.

(3) Expansion with Rules Q<=>X

We choose a lower confidence of 0.1 and support of 0.0001 here, since there are not many such rules with high confidences in both directions. The rules for the query of No. 9 are

森林 (forest) <=> 森林资源 (forest resource), 生态 (ecology), 林区 (forest area),林业 (forestry);

火灾 (fire disaster) <=> 消防 (fire control/prevention/fighter), 大火 (conflagration), 火势 (fire impetus), 防火 (fire prevention).

After expansion, the average precision was promoted from 0.45 to 0.78 or 73.33%. This method expands all the 10 queries. Improvement of the overall average precision is 14.77% (from 0.44 to 0.505). We get a significant improvement in performance with the query of No. 3, but there are three queries with which a decreased performance is found as well. Moreover, this method shows inefficient with rank 1 and rank 9, and beneficial with other ranks. This is the worst among the three expansion methods. The performance of 10 ranks is shown in Figure 5.

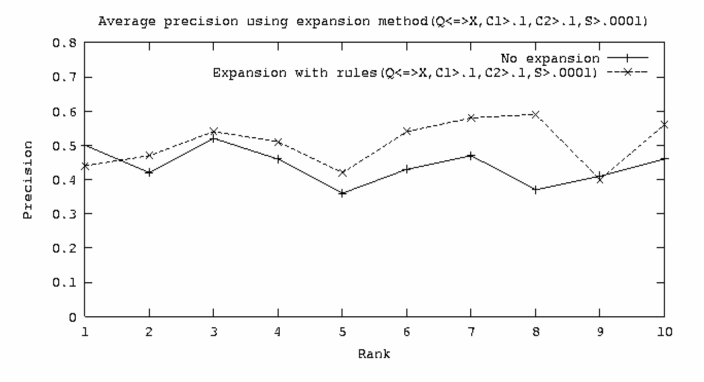

Figure 5: The average precision for each rank with rules Q<=>X.

IV. Conclusions and Future Work

Recognizing the importance of Chinese information, we have developed a system for automatically generating Chinese-crossed-query expansions with CTAR mining techniques. The most feature of our system is to segment Chinese information into clusters such that each cluster contains those words that have a same character. We have demonstrated that our method can improve the effectiveness of retrieval over the corpus from the Xinhua News Agency. It is quite natural and reasonable to apply correlations among terms to the query expansion. Therefore, we believe that the idea in this paper is promising.

In the future work, we still need to study the thresholds determination for the actionable CTAR mining. Since there are many expansion methods, we should consider how to integrate these methods.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments on the first version of this paper. These comments have contributed to a vast improvement of this paper.
Also, we would like to thank Jie Wei and Dr. Shichao Zhang for their cooperative and discussion work for this paper.

References

[1] Efthimis N. Efthimiadis. Query expansion. In: Martha E. Williams edited, *Annual Review of Information Scienc and Technology* (ARIST), Volume 31, pages 121-187, 1996.

[2] Mathias Géry and M. Hatem Haddad. Knowledge discovery for automatic query expansion on the World-Wide Web. In: Proceedings of Advances in Conceptual Modeling: ER '99 Workshops, Lecture Notes in Computer Science 1727, Springer, pages 334-347, Paris, France, November 15-18, 1999.

[3] Hatem Haddad, J.P. Chevallet, M.F. Bruandet. Relations between Terms Discovered by Association Rules. In: Proceedings of the 4th European Conference on Principles and Practices of Knowledge Discovery in Databases PKDD'2000, Workshop on Machine Learning and Textual Information Access, Lyon France, September 12, 2000.

[4] Jie Wei, Stéphane Bressan, Beng Chin Ooi. Mining Term Association Rules for Automatic Global Query Expansion: Methodology and Preliminary Results. Proceedings of First International Conference on Web Information Systems Engineering (WISE'00)-Volume 1

[5] R. Feldman, M. Fresko, Y. Kinar, Y. Lindell, O. Liphstat, M. Rajman, Y. Schler, O. Zamir. Text Mining at the Term Level. In Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), pages 65-73, Nantes, France. September 1998.

[6] Maria-Luiza Antonie, Osmar R. Zaïane. Text Document Categorization by Term Association. In Proceedings of 2002 IEEE International Conference on Data Mining (ICDM'02), pages 19-26, Maebashi City, Japan, December 09 - 12, 2002.

[7] Jie Wei, Zhenxing Qin, Stéphane Bressan, Beng Chin Ooi. Mining Term Association Rules for Automatic Global Query Expansion: A Case Study with Topic 202 from TREC4. In Proceedings of Americas Conference on Information Systems 2000.

[8] Xiaowei Yan, Chengqi Zhang, Shichao Zhang: Identifying Frequent Terms in Text Databases by Association Semantics. In: Proceedings of 2003 International Symposium on Information Technology (ITCC 2003), 28-30 April 2003, Las Vegas, NV, USA. IEEE Computer Society 2003: 672-675

[9] Chengqi Zhang and Shichao Zhang, Association Rules Mining: Models and Algorithms. Springer-Verlag Publishers in Lecture Notes on Computer Science, Volume 2307, p. 243, 2002.

[10] Shichao Zhang, Chengqi Zhang and Xiaowei Yan, PostMining: Maintenance of Association Rules by Weighting. Information Systems, Volume 28, Issue 7, October 2003: 691-707.

[11] “WordNet - a Lexical Database for English”, www.cogsci.princeton.edu/~wn/

[12] http://www.creadersnet.com/newsPool/.