Text Mining Based Query Expansion for Chinese IR

Zhihan Li Yue Xu Shlomo Geva

Faculty of Information Technology
Queensland University of Technology
Brisbane, QLD 4001, Australia
Z7.li@student.qut.edu.au {yue.xu,s.geva}@qut.edu.au

Abstract

Query expansion has long been suggested as a technique for dealing with word mismatch problem in information retrieval. In this paper, we describe a novel query expansion method which incorporates text mining techniques into query expansion for improving Chinese information retrieval performance. Unlike most of the existing query expansion strategies which generally select indexing terms from the top N retrieved documents and use them to expand the query, in our proposed method, we apply text mining techniques to find patterns from the retrieved documents which contain relevant terms to the query terms, then use these relevant terms which can be indexing terms or indexing term patterns to expand the query. The experiment with NTCIR-5 collection shows apparent improvement in both precision and recall.

1 Introduction

The amazing growing speed in the number of Chinese Internet users indicates that building Chinese information retrieval systems is in great demand. This paper presents a method aimed at improving the performance of Chinese document retrieval.

Unlike English text in which sentences are sequences of words delimited by white spaces, in Chinese text, sentences are represented as strings of Chinese characters without separating spaces between words. For Chinese information retrieval, the query is usually a set of Chinese words rather than a sequence of Chinese characters. For character based Chinese information retrieval, since the texts are not segmented, the retrieved documents which contain the character sequence of the query may not be relevant to the query since they may not contain the words in the query. Therefore, the quality of character based Chinese information retrieval is not satisfactory. On the other hand, a study has shown that the relationship between segmentation and retrieval is in fact non-monotonic, that is, high precision of segmentation alone may not improve retrieval performance (Peng, Huang, Schuurmans and Cercone 2002). In this paper we propose a Chinese information retrieval model which combines character based retrieval and word based ranking to achieve better performance.

Another critical problem is that the original query may not represent adequately what a user wants. By appending additional terms to the original query, Query Expansion attempts to construct a richer expression to better represent the user's information need. Pseudo relevance feedback is a popular technique for query expansion. The basic idea is to automatically extract additional relevant terms from the top ranked documents in the initial result list. These terms are added to the original query, and the extended query is executed with the expectation of improved performance. In this paper, we propose a new approach to improving the performance of Chinese document retrieval by expanding queries with highly correlated segmented words, generated by using text mining techniques.

The remainder of this paper is structured as follow. Section 2 briefly reviews some related work. In Section 3, we describe our retrieval model, and then present the text mining based query expansion method in Section 4. Section 5 presents experimental results and Section 6 concludes the paper.

2 Related Work

Unlike English text in which sentences consist of words delimited by white spaces, in Chinese text, sentences are represented as strings of Chinese characters without delimiters. Therefore, Chinese word segmentation is the first phase in Chinese language processing and has been widely studied for many years (Gao and Li 2005; Xue 2003; Sproat and Shih 2002; Wang, Liu and Qin 2006). Both Chinese characters and words can be used as the indexing units

for Chinese IR. Several approaches have shown that single character indexing can produce good results, but word and bi-gram indexing can achieve slightly better performance. This however incurs greater time and space complexity with limited performance improvement (Sproat and Shih 2002; Li 1999; Kwok 1997; Peng, Huang, Schuurmans and Cercone 2002). In this paper, we propose a ranking method that combines character indexing and segmented word indexing to re-rank retrieved documents and promote relevant documents to higher positions.

Pseudo-relevance feedback is an important query expansion technique for improving IR performance (Qiu and Frei 1993; Sun, Ong and Chua 2006; Robertson and Jones 1976). The basic insight which motivates pseudo relevance feedback is that often the top of the initially ranked list of results contains a relatively high proportion of relevant documents. The conjecture is that despite the presence of some irrelevant documents, these retrieved documents might still be used to identify relevant terms that co-occur in the relevant documents. These terms are then used to modify the original query and better reflect the user's information needs. With the expanded query, a second retrieval round is performed and the returned result is expected to contain more relevant documents which have been missed in the first retrieval round. For pseudo relevance feedback query expansion, the most important task is to find the terms from the retrieved documents that are considered relevant to the query. Therefore, relevant term selection is crucial in pseudo relevance feedback query expansion. The standard criteria for selecting relevant terms have been proposed using tf/idf in vector space model (Rocchio 1997) and probabilistic model (Robertson and Jones 1976). Query length has been considered in (Kwok, Grunfeld and Chan 2000) for weighting expansion terms and some linguistic features also have been tried in (Smeaton and Rijsbergen 1983). We are proposing to use text mining techniques to find the relevant terms.

Data Mining is about analyzing data and finding hidden patterns using automatic or semi-automatic means. Text mining is a research field of data mining which refers to the process of deriving high quality patterns and trends from text. We are proposing to apply text mining techniques to finding frequent patterns in the retrieved documents in the first retrieval round which contain query terms. These patterns provide us with the candidate sequences to find more terms which are relevant to the original query.

The application of text mining to information retrieval may improve precision and recall.

3 Retrieval Model

In general, character indexing based IR can retrieve most of the relevant documents as long as they contain the query terms (the query terms are sequences of Chinese characters but not necessarily sequences of segmented words in the retrieved documents since the documents are not segmented). However, the retrieval performance is not necessarily good. This is because many irrelevant documents are highly ranked due to high query term frequency corresponding to instances of the query term sequences which are actually not valid words but rather correspond to what would be incorrect word segmentation. On the other hand, the word indexing based IR can apply better ranking and therefore achieve somewhat better performance than that of character indexing based IR. The improvement is limited since some relevant documents may not contain the query terms as segmented words and thus won't be retrieved. In this section, we describe our approach to Chinese information retrieval in which, we firstly create two indexing tables from the data collection, a Chinese character indexing table and a segmented Chinese word indexing table; secondly, the relevant documents are retrieved based on the character indexing, then the retrieved documents are ranked by a method proposed in this paper that ranks the retrieved documents based on both character indexing and word indexing.

3.1 Indexing and Retrieval Model

The first task is word segmentation. We used an online program provided by the Institute of Information science, Academia Sinica in TaiWan to segment the documents. The segmentation precision of the system is approximately 95% as reported in (Ma and Chen 2002), and most importantly, this system not only accomplishes word segmentation, but also incorporates POS (part of speech) annotation information into the segmented documents which is very important for this research since we are therefore able to utilize the POS information to improve the efficiency and effectiveness of Chinese IR based on word indexing. For example, for the following original Chinese text:

在俄羅斯的救援行動失敗之後，一艘英國迷你潛艇正緊急馳援途中，預計 19 日晚上 7 時台北時間 19 日晚 23 時可抵達科斯克號核子動力潛艇的沉沒現場，展開救援工作。

We can get the following segmented text:

Figure 1. Chinese Word Segmentation Example

The left part of figure 1 is the segmented document and the right part lists all words in the segmented document and each word is associated with a POS tag immediately after the word.

We built two indexing tables, one is character based and the other is segmented word based. In earlier research (Lu, Xu and Geva 2007), we have developed a character based indexing system to perform character based retrieval and ranking. In this paper, we directly use the existing retrieval model, which is the traditional Boolean model, a simple model based on set theory and Boolean algebra, to perform document retrieval.

3.2 Ranking Method

In our previous research, we used the following ranking model to calculate the ranking value of a retrieved document to determine the top N relevant documents:

$$d_{rank} = n^5 \sum_{i=1}^{m} tf_i^c \times idf_i^c \quad (1)$$

Here, *m* is the number of query terms, *n* is the number of distinct query terms that appear in the document as character sequences (not necessarily segmented words). tf_i^c is the frequency of the i^{th} term in the document and idf_i^c is the inverse document frequency of the i^{th} term in the collection. The equation can ensure two things: firstly, the more distinct query terms are matched in a document, the higher the rank of the document. For example, a document that contains four distinct query terms will almost always have higher rank than a document that contains three distinct query terms, regardless of the query terms frequency in the document. Secondly, when documents contain a similar number of distinct terms, the score of a document will be determined by the sum of query terms' *tf-idf* value, as in traditional information retrieval.

In this paper, we use the following equation to calculate documents' ranking scores, which is simply the average of character based ranking and word based ranking:

$$d_{rank}^{c+w} = \frac{n_c^5 \sum_{i=1}^{m} tf_i^c \times idf_i^c + n_w^5 \sum_{k=1}^{m} tf_i^w \times idf_i^w}{2} \quad (2)$$

Where n_c is the number of distinct query terms which appear in the document as character sequences (not necessarily segmented words). tf_i^c and idf_i^c are the same as in Equation (1), tf_i^w is the frequency of the i^{th} term in the document as a segmented word, and idf_i^w is the inverse document frequency of the i^{th} term in the collection as a segmented word, and n_w is the number of query terms which appear in the document as a segmented word.

4 Query Expansion

It has been recognized that user queries consisting of a few terms are inadequate to fully reflect users' information needs and lead to a poor coverage of the relevant documents. Query expansion is the technique widely used to deal with this problem. In this section, we describe a new method that applies text mining techniques to find terms from the retrieved documents that are highly correlated to the query, and then performs a second retrieval using the expanded query with these relevant terms. In the first part of this section, we introduce the method to generate a set of candidate relevant terms from the retrieved documents. Then, in the second part of this section, we introduce a method to select the most relevant terms to expand the original query.

In this stage, the top N retrieved documents from the first retrieval round are converted to a set of transactions which are used to mine frequent patterns using text mining methods such as FP-Tree method (Han, Pei and Yin 2000; Zou, Chu, Johnson and Chiu 2001; Agrawal and Srikant 1994). Query terms are usually nouns. So, it is reasonable to only extract patterns of nouns rather than patterns of all

words in the retrieved documents. Therefore, from the retrieved documents we first eliminate all non-noun words based on POS tag information, then construct a collection of transactions each of which consists of the nouns in a sentence from a retrieved document. Thus, all sentences in the retrieved documents are included in the transaction collection if they contain nouns. A sentence will be excluded from the transaction collection if it does not contain nouns. For example, we extracted 18 unique nouns from the example in Figure 1. The 18 nouns form a vector with size 18:

俄羅斯(Russia), 行動 (Action), 英國 (England), 潛艇 (Submarine), 途中 (Way), 19 日 (19th), 晚上 (Evening), 7 時 (7PM), 台北 (TaiBei), 時間 (Time), 晚 (Night), 23 時 (23PM), 科斯 (Kesi), 號 (Name), 核子 (Nucleon), 動力 (Power), 現場 (Scene), 工作(Work).

Each transaction created from one sentence in a retrieved document is a vector of 1s and 0s, each element of the transaction corresponds to a noun with value 1 indicating the corresponding noun appearing in the sentence and otherwise 0. The number of transactions is the number of sentences in the retrieved top N documents. The size of the transaction is the number of unique nouns in the top N documents.

The collection generated from the example is shown in Table 1:

sentence	*Transaction*
在俄羅斯的救援行動失敗之後	110000000000000000
一艘英國迷你潛艇正緊急馳援途中	001110000000000000
預計 19 日晚上 7 時台北時間 19 日晚 23 時可抵達科斯克號核子動力潛艇的沉沒現場	000101111111111110
展開救援工作	000000000000000001

Table 1. Example collection of transactions

Any pattern mining algorithm can be used to generate frequent patterns from the constructed collection mentioned above. In this paper, we choose the popular used FP-Tree algorithm to perform the pattern mining task. Let Q={q_1, …, q_m} be the query containing m query terms and $FP = \{p_1, p_2, \cdots, p_n\}$ be the set of mined frequent patterns, $Supp(p_k)$ be the support value of pattern p_k. For a query term q_i, the set of patterns which contain the query term is denoted as $FP_i = \{p_k | p_k \in FP, q_i \in p_k\}$. For a pattern p_j and a query term q_i, the set of patterns which contain both p_j and q_i is denoted as $FP_{ij} = \{p_k | p_k \in FP, q_i \in p_k, p_j \in p_k\}$, and the set of patterns which contain p_j is denoted as $FP_j = \{p_k | p_k \in FP, p_j \in p_k\}$.

The following equation is proposed to calculate the relevancy between a query term q_i and a pattern p_j which is denoted as $R_{(q_i,p_j)}$:

$$R_{(q_i,p_j)} = W_1 \frac{N_{ij}}{N_i} + W_2 \frac{\sum_{p_k \in FP_{ij}} Supp(p_k)}{N_{ij}} + W_3 \frac{\sum_{p_k \in FP_j} Supp(p_k)}{N_j} \quad (3)$$

Where $N_{ij} = |FP_{ij}|$, $N_i = |FP_i|$, $N_j = |FP_j|$, which are the number of frequent patterns. The first part in Equation (3) measures the ratio between the patterns which contain both the query term q_i and the pattern p_j. The higher the ratio is, the more the pattern p_j is relevant to the query q_i. The second part in the equation is the average support value of the patterns containing both the query term q_i and the pattern p_j. A high average support indicates that the co-occurrence of the query term q_i and the pattern p_j is high which in turn indicates that q_i and p_j are highly correlated with each other. Similarly, the third part is the average support value of the patterns containing only the pattern p_j. A high average support of pattern p_j means that p_j is a popular pattern in the retrieved documents and may have a high relevance to the query term. Equation (3) is a linear combination of the three parts with W_1, W_2 and W_3 as coefficients which can be controlled by users to adjust the weights of the three parts.

For the whole query, the relevancy value R_{p_j} of pattern p_j is calculated as follows:

$$R_{p_j} = 100 \times \sum_{q_i \in Q} R_{(q_i,p_j)} \quad (4)$$

All these patterns are ranked by relevancy values. Based on the relevancy value R_{p_j}, we select the top 3 patterns and use the words in the patterns to expand the original query.

5 Experimental Results

We select NTCIR-5 Chinese corpus as our dataset, which includes 434,882 documents in traditional Chinese. We built two indexing tables, one contain indexes of all Chinese characters appearing in the documents, and the other contains all Chinese words with POS information produced by the online segmentation program developed by the Institute of Information science, Academia Sinica in TaiWan. Fifty queries are used and each query is a simple description which consists of one or several terms. We use the average precision (denoted as P in Table 2) and average recall (denoted as R) of the top 10, 15, 20, 30 and 100 retrieved documents to evaluate the performance of the proposed Chinese IR model with the text mining based query expansion (denoted as QE(TM)) by comparing with the character based model without query expansion (denoted as C) , the word-character based model without query expansion (denoted as W-C), and the proposed Chinese IR model with the popularly used standard query expansion method Rocchio (denoted as QE(R)). In the experiments, W1, W2, W3 are set to 0.8, 0.18, 0.02, respectively. Additionally, we set the support value as 0.1 for using FP-Tree method to mine frequent patterns. The experiment results are given in Table 2.

TOP N	C		W-C		QE (R)		QE(TM)	
	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)	P (%)	R (%)
10	39.7	19.6	39.9	20.1	41.6	23.2	49.2	26.8
15	35.2	26.1	35.5	26.5	39.7	28.9	44.1	34.0
20	32.7	30.9	32.9	31.3	35.9	32.1	39.2	38.7
30	27.5	36.7	27.5	35.7	30.2	38.8	33.0	44.5
100	14.1	53.8	14.6	53.9	15.1	60.4	16.3	64.7
Ave	29.9	33.4	30.1	33.5	32.5	36.8	36.4	41.7

Table 2 Precision and Recall

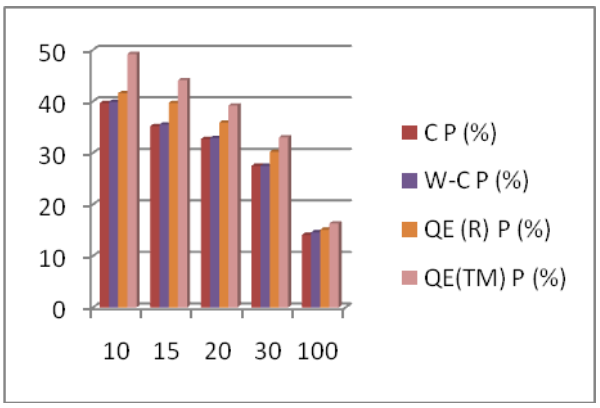

Figure 2 Precision

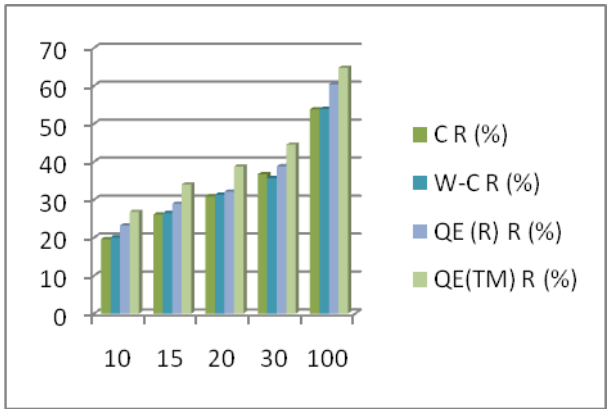

Figure 3 Recall

Table 2 shows the precision and recall of the four different retrieval approaches. From Table 2, we can see that the performance is improved slightly from the character based model to the word-character based model, the precision is improved by 0.2% on average from 29.9% to 30.1% and the recall is improved by 0.1% on average from 33.4% to 33.5%. With the Rocchio standard query expansion, we achieved a little more improvement, the precision is improved by 1.4% on average from 30.1% to 32.5% and the recall is improved by 3.3% from 33.5% to36.8%. However, with our text mining based query expansion method, we achieved much larger improvements; precision is improved by 6.3% on average from 30.1% to 36.4% and the recall is improved by 8.2% from 33.5% to 41.7%.

6 Conclusion

In this paper, we proposed an approach to improve the performance of Chinese information retrieval. The approach includes two aspects, retrieval based on segmented words and text query expansion using text mining techniques. The experiment results show that Chinese word segmentation does not bring significant improvement to Chinese informa-

tion retrieval. However, the proposed text mining based query expansion method can effectively improve the performance of the Chinese IR and the improvement is much greater than that achieved by using the standard query expansion method Rocchio.

References

Chengye Lu, Yue Xu, Shlomo Geva. 2007. *Translation disambiguation in web-based translation extraction for English-Chinese CLIR*. Proceedings of the 2007 ACM symposium on Applied computing, 819-823.

F. Peng, X. Huang, D. Schuurmans, and N. Cercone. 2002. *Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR*, Proceedings of the 19th international conference on Computational linguistics, 1-7.

J. Han, J. Pei and Y. Yin. 2000. *Mining Frequent Patterns without Candidate Generation*. Proc of the 2000 ACM International Conference on Management of Data, Dallas, TX, 2000, 3-12.

Jianfeng Gao and Mu Li. 2005. *Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach*. Computational Linguistics, MIT. 531-574, Vol 31, Issue 4.

K. L. Kwok. 1997. *Comparing representations in Chinese information retrieval*. Proc. Of the ACM SIGIR97, 34-41.

Kwok, K.L, Grunfeld, L., Chan, K. 2000. *THREC-8 adhoc, query and filtering track experiments using PIRCS*, In TREC10.

Nianwen Xue. 2003. *Chinese Word Segmentation as Character Tagging*. Computational Linguistics and Chinese Language Processing, Vol 8, No 1, Pages 29-48.

P. Li. 1999. *Research on improvement of single Chinese character indexing method*. Journal of the China Society for Scientific and Technical Information, 18(5).

Q. Zou, W. Chu, D. Johnson and H. Chiu. 2001. *A Pattern Decomposition (PD) Algorithm for Finding all Frequent Patterns in Large Datasets.* Proc of the 2001 IEEE International Conference on Data Mining (ICDM01), San Jose, California, 673-674.

Qiu Y. and Frei H. 1993. *Concept based query expansion*. In Proceedings of SIGIR 1993, pp. 160-169.

R. Agrawal and R. Srikant. 1994. *Fast algorithms for mining association rules*. Proc. Of the 1994 International Conference on Very Large Data Bases, Santiago, Chile, 487-499.

Renxu Sun, Chai-Huat Ong, Tat-Seng Chua. 2006. *Mining dependency relations for query expansion in passage retrieval.* Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, Pages: 382 – 389.

Richard Sproat and Chilin Shih. 2002. *Corpus-Based Methods in Chinese Morphology and Phonology*.

Robertson, S.E. and K. Sparck Jones. 1976. *Relevance Weighting of Search Terms*. Journal of the American Society for Information Science, 27(3): 129-146.

Rocchio, J. 1997. *Relevance feedback in information retrieval. In the Smart Retrieval System: Experiment in Automatic Document Processing*, Pages 313-323.

Smeaton, A. F. and Van Rijsbergen, C. J. 1983. *The retrieval effects of query expansion on a feedback document retrieval system*.Computer Journal, 26(3): 239-246.

Wei-Yun Ma, Keh-Jiann Chen. 2002. *Introduction to CKIP Chinese Word Segmentation System for the First International Chinese Word Segmentation Bakeoff.* Institute of Information science, Academia Sinica.

Xinjing Wang, Wen Liu and Yong Qin. 2006. *A Search-based Chinese Word Segmentation Method*. Proceedings of the 16th international conference on World Wide Web.