Query Based Chinese Phrase Extraction for Site Search

Jingfang Xu, Shaozhi Ye, and Xing Li

Department of Electronic Engineering, Tsinghua University
Beijing 100084, P.R.China
xjf02@mails.tsinghua.edu.cn, ys@compass.net.edu.cn, xing@cernet.edu.cn

Abstract. Word segmentation(WS) is one of the major issues of information processing in character-based languages, for there are no explicit word boundaries in these languages. Moreover, a combination of multiple continuous words, a phrase, is usually a minimum meaningful unit. Although much work has been done on WS, in site web search, little has been explored to mine site-specific knowledge from user query log for both more accurate WS and better retrieval performance. This paper proposes a novel, statistics-based method to extract phrases based on user query log. The extracted phrases, combined with a general, static dictionary, construct a dynamic, site-specific dictionary. According to the dictionary, web documents are segmented into phrases and words, which are kept as separate index terms to build phrase enhanced index for site search. The experiment result shows that our approach greatly improves the retrieval performance. It also helps to detect many out-of-vocabulary words, such as site-specific phrases, newly created words and names of people and locations, which are difficult to process with a general, static dictionary.

1 Introduction

Information retrieval(IR) systems select relevant documents by matching index terms with the query. The selection of index terms affects both precision and efficiency of retrieval systems. An ideal indexing term should be a meaningful unit, expressing the concept documents contain and the information user need[1]. Words, the minimum meaningful units, are selected as index terms in most IR systems. Phrases, comprising several continuous words, are usually also inseparable combinations, i.e., they might be misunderstood when broken into words. Therefore, phrases should also be kept as index terms. In the retrieval of alphabet-based languages, such as English and French, it has been proved that choosing both words and phrases as indexing terms is more efficient than indexing words only[2]. Unlike words, phrases have no explicit separator, thus phrase extraction turns out to be a challenge. It becomes more complex when processing character-based languages, such as Chinese, which have neither word boundary nor phrase separator. To deal with these languages, word segmentation has to be done before phrase extraction.

X. Zhou et al. (Eds.): WISE 2004, LNCS 3306, pp. 125–134, 2004.

The basic approaches of word segmentation in character-based languages can be partitioned into two categories: statistic-based and dictionary-based[3]. Statistic-based approaches make use of statistical properties, such as frequencies of characters and character sequences in the corpus[4]. Mutual information(MI) is usually employed by these approaches[4][5]. Dictionary-based approaches use a dictionary to identify words. When matched with a word in the dictionary, a sequence of characters will be extracted as a word. For match approaches, there are maximum match, minimum match and hybrid approach. The maximum match approach can be further divided into forward maximum match approach and backward maximum match approach. Ideally, the dictionary-based approach can detect all the words if the dictionary is complete. In practice, however, a static dictionary that contains all possible words is unfeasible, costly and unnecessary[6]. A corpus usually contains only part of words in the dictionary, and on the other hand, a static dictionary is lack of many out-of-vocabulary words, e.g., site-specific words, newly created words and names of people and locations. A dynamic, topic-specific dictionary of a corpus will alleviate the problem by eliminating irrelevant words and providing special words.

Phrase extraction is similar to word segmentation. There is already some work on the statistics-based phrase extraction[7][8]. Previous statistics-based phrase extraction approaches consider all possible phrases in the corpus as potential phrase candidates and calculate the statistical information of all these phrases, which is costly and time consuming. Moreover, the extracted phrases are often non-meaningful or never concerned by users. In this paper we propose a method applying MI to phrase extraction based on user query log. We obtain the phrase candidates only from user queries and calculate their MI scores in both document space and query space. The experiment result shows the efficiency of our method and the extracted phrases are really needed by users.

To evaluate our phrase extraction algorithm, we build a phrase enhanced index, which is supposed to improve both precision and efficiency of retrieval. Phrase index terms provide more accurate information, thus help to select more relevant results. On the other hand, indexing phrases accelerate phrase search. When searching a phrase, e.g., "AB", if "AB" is not an index term, all the documents that contain both A and B have to be chosen to check whether "A" is followed by "B". It will be much faster if the phrase "AB" is an index term.

In this paper, we choose a Chinese web site search engine for our experiment platform and use 69,680 web documents and 508,464 query requests as training set. First, with a static, general dictionary, we segment the training set into words and extract phrases based on MI. Then we combine these phrases with a general dictionary to construct a dynamic, site-specific dictionary. This dictionary is used to parse all web documents into words and phrases and finally build a phrase enhanced index. The experiment result shows that when queries of testing set match the general word only index, the average number of index terms per query hits is 2.61 and with our enhanced index, it can be reduced to 1.93. Although the results presented in this paper are all based on Chinese documents, our

method can be easily adopted in other character-based languages with just a little modification of character encoding.

The rest of this paper is organized as follows. First we propose our phrase extraction method in Section 2 and analyze the experiment dataset in Section 3. Then experiment results and evaluation are presented in Section 4. Finally we review the previous work in Section 5 and conclude this paper with Section 6.

2 Mutual Information and Phrase Extraction

Mutual information(MI) can be used to measure the coherence of the adjacent characters and is applied widely in statistic-based WS, where the adjacent characters with high MI score are identified as a word. In our approach, similarly, we identify the adjacent words as a phrase if its MI score is higher than a predefined threshold.

Consider a string of words "$...c_0c_1c_2c_3...$", the MI of words c_1 and c_2 is computed by Equation 1:

$$MI(c_1c_2) = \log_2 \frac{p(c_1c_2)}{p(c_1)p(c_2)} \tag{1}$$

Where $p(c_1c_2)$ is the occurrence probability of the words sequence "c_1c_2", which is estimated by the times that c_1 is followed by c_2, normalized by N, the total number of words in the corpus. $p(c)$ is the probability of word c, which is estimated by the total occurrences of the word c normalized by N. Therefore, Equation 1 is represented as Equation 2:

$$MI(c_1c_2) = \log_2 \left(\frac{\frac{freq(c_1c_2)}{N}}{\frac{freq(c_1)}{N}\frac{freq(c_2)}{N}} \right) = \log_2 \left(N \frac{freq(c_1c_2)}{freq(c_1)freq(c_2)} \right) \tag{2}$$

The previous phrase extraction methods select all the potential phrases in documents, while we only consider the potential candidates which have appeared in the user query log. More precisely, three methods to calculate MI are compared in this paper.

2.1 MI in Query Space

The query space is constructed by all the user queries. In query space we tend to extract the phrases which interest users while their separated constituent words do not. Here N in Equation 2 is the total number of words in query log, and the $freq(c)$ is estimated by the times of word c appears in the log.

2.2 MI in Document Space

The document space is constructed by all the web documents. In this space, we consider the occurrences of phrase candidates in web documents. For all the candidates, we compute their MI scores in the document space. Here N is the total number of words in web documents and the $freq(c)$ is estimated by the times of word c appears in the web documents.

2.3 MI in Hybrid Space

Method 1 extracts the phrases that usually queried by users and method 2 extracts the phrases that frequently appear in the documents. This method integrates them, i.e., extracts the phrases that occur frequently in both user queries and web documents. We first compute MI scores of the candidates in query space, and discard the phrases whose MI scores are lower than a threshold. Then, we compute MI scores of the phrases extracted from query space in document space to select the high MI score phrases.

Combined with the general dictionary, the extracted phrases construct a site-specific dictionary. As we track the dynamic query log, the site-specific dictionary with extracted phrases is also dynamic.

3 Dataset Analysis

We choose a Chinese web site, the homepage of *China Education and Research Network(CERNET)*[1], and its site search engine as our experiment platform. Table 1 summarizes the web documents of this site and training query log from its site search engine. Forward maximum match method is used to segment 69,680 web documents and 508,464 Chinese requests into words, according to a general dictionary from *On-Line Chinese Tools*[2] which contains 119,804 unique Chinese words.

Table 1. Summary of Web Documents and Query Log

Web Documents	69,890
Words in Documents	40,889,563
Unique Words in Documents	82,186
Log Days	Aug. 13, 2003 - Jan. 31, 2004
Total Queries	627,920
Chinese Queries	508,464
Unique Queries	82,078
Words in Queries	1,335,151
Unique words in Queries	13,808

The analysis of the training set shows that:

1. Neither web documents nor queries contain all the words in the general dictionary. Web documents contain 69% and the query log contains only 9%, which means that many words in the general dictionary never appear in web documents or queries.

[1] http://www.edu.cn
[2] http://www.mandarintools.com/

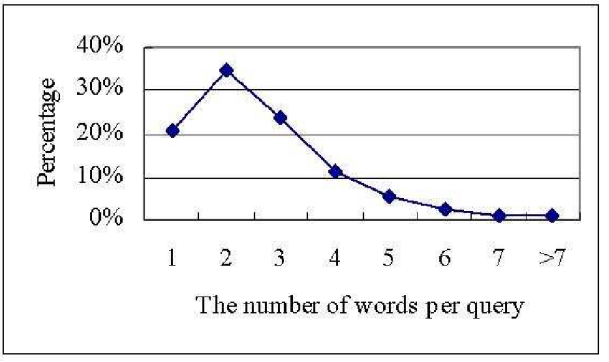

Fig. 1. Words Per Query in the Log

2. Most queries contain more than one word. As shown in Figure 1, only 20.93% queries contain just one word and 34.67% queries are made up of two words. Searching multiple words is more complex and time consuming than searching a single word[9]. So indexing phrases will greatly improve retrieval performance. Moreover, many query phrases are site-specific or newly created, which can not be dealt with a general, static dictionary. These phrases may be related to many other documents when broken into words, which hurts both precision and efficiency. Thus these specific query phrases should be extracted and indexed directly. In this paper, the notion phrase enhanced index is used for indexing both words and phrases, comparing with the original index which indexes only words.
3. The distribution of words in web documents differs from that in query log. Removed stop words, the top 10 most frequent words in the web documents are totally different with the top 10 most frequently queried ones, as shown in Figure 2. There are two possible reasons. First, many words in the web document are not interested by users. Second, web documents and queries may use different words to describe the same thing. The words in web documents are usually written and official, while those in user query are mostly spoken and informal. Moreover, users may also use different queries to search just the same information. For an IR system, it may be not interested in what the documents contain, but it may concern what the user want to search. Those extracted phrases, never searched by user, are useless to the system and should not be indexed as an index term. So it is unnecessary to extract phrases through full document.

4 Experiment Results and Evaluation

First of all, all the web documents and Chinese queries are segmented into words according to the general dictionary. 400,477 queries which occur in the log and

Most Frequent Words in Document Space	Most Frequent Words in Query Space
教育(Education)	公务员(Official)
中国 (China)	成人高考 (Adult College Entrance Exam)
稿件 (Contribution)	专升本 (Upgrade from Junior College Student to University Student)
转载(Outside Post)	成绩查询(Grade Enquiry)
科研(Research)	论文(Paper)
大学(University)	公务员考试(Official Examination)
评论(Comments)	成绩(Grade)
问题(Question)	简历(Resume)
注明 (Indicate)	计算机等级考试 (Computer Rank Exam)
学生(Student)	答案(Answer)

Fig. 2. Most Frequent Words in Documents and Queries

contain more than one word are selected as phrase candidates. Their MI scores in three spaces are computed separately. In hybrid space, we first eliminate the phrases whose MI scores in query space are lower than 13[3], and then compute MI scores of the rest candidates in document space. Figure 3 lists top 15 phrases with highest MI scores in each space. In each space we extract the top 8,000 MI score phrases and combine them with the same general dictionary to construct a dynamic dictionary for the corresponding space.

To evaluate three phrase extraction methods, phrase enhanced indices are built which index both words and phrases in the corresponding dictionaries. The general dictionary and three dynamic dictionaries are used to segment web documents and queries individually. Then we build a general word only index and three phrase enhanced indices. The notions query index, document index and hybrid index are used for phrase enhanced indices built with the query space, document space and hybrid space dictionaries respectively.

Table 2. Summary of Test Queries

Log Days	Feb.1 - Feb.29,2004
Total Queries	75,499
Chinese Queries	60,235
Unique Queries	24,083

[3] We check the phrases with high MI scores manually and select this threshold.

Query Space	Document Space	Hybrid Space
课件(Courseware)	新婚姻(New Marriage)	新婚姻(New Marriage)
雅思 (International English Language Test)	民族团结 (National Unity)	民族团结 (National Unity)
三个代表 (Three Representatives)	中国青年政治学院 (China Youth University for Political Sciences)	通货紧缩 (Deflation)
八级 (Eighth-band Test for English Majors)	通货紧缩 (Deflation)	实话实说 (Tell the Truth)
申论(State)	实话实说(Tell the Truth)	中国青年政治学院 (China Youth University for Political Sciences)
亿唐 (Etang: Name of a Website)	新世纪 (New Century)	珠江三角洲 (Delta of Zhujiang River)
说课(Teaching)	未就业(Not Employed)	新世纪(New Century)
六级(College English Test 6)	乙型肝炎(Hepatitis B)	未就业(Not Employed)
自我鉴定 (Appraise Oneself)	完形填空 (Cloze)	给水排水 (Supply Water and Drain off Water)
英语学习(English Study)	近世代数(Modern Algebra)	乙型肝炎(Hepatitis B)
特长生(Students Having Specialty)	神州数码(Digital China)	完形填空(Cloze)
北外(Being Foreign Studies University)	线性代数(Linear Algebra)	近世代数(Modern Algebra)
课改(course reforming)	期末考试(Final Exam)	化学方程式(Chemical Equation)
非典(SARS)	居里夫人(Mrs. Curie)	神州数码(Digital China)
查分 (Grade Enquiry)	同等学力 (Same Educational Level)	敦煌莫高窟 (Mogao Grottoes of Dunhuang)

Fig. 3. Top 15 MI Score Phrases in Each Space

The user queries of CERNET from Feb. 1, 2004 to Feb. 29, 2004 are used as our testing data. Table 2 summarizes the testing queries. These testing queries are matched by four indices, general index, query index, document index and hybrid index. The hit ratios are shown in Table 3 and Figure 4.

When retrieval, if queries hit fewer index terms, there will be fewer index operations. For example, if the query is segmented into just one index term, we can directly return the record of this index term. While if the query is segmented into three index terms, the records of these three index terms have to be retrieved and compared to generate the final results for user. Multi-index term accesses will be prohibitively expensive especially when the index is large, thus reducing index term accesses is vital to alleviate the overhead of retrieval. Table 3 shows that all the three phrase enhanced indices outperform the original word only index in one index term hit ratio, which means a direct hit and no operation for multi index terms comparison. The hybrid index gets a 22.93% higher direct hit ratio than the general index. The average number of index terms per query is 2.61 with the general index and it is reduced to 1.93 with the hybrid index.

Table 3. Query Hit Ratios of Different Indices

Index Terms Per Query	General Index	Query Index	Document Index	Hybrid Index
1(direct hit)	20.93%	32.26%	43.55%	45.86%
2	34.67%	34.57%	30.46%	30.91%
3	23.49%	19.24%	15.74%	13.91%
4	11.10%	8.16%	6.07%	5.64%
5	5.48%	3.21%	2.34%	2.58%
6	2.35%	1.39%	0.96%	0.08%
7	1.03%	0.55%	0.41%	0.46%
>7	0.95%	0.68%	0.46%	0.56%
Average	2.61	2.25	2.00	1.93

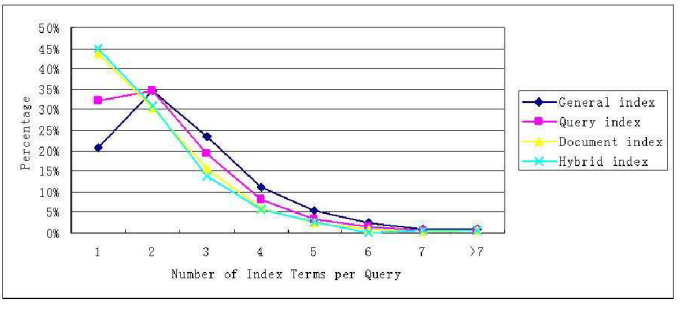

Fig. 4. Index Terms Hit by Queries to Different Indices

The results show that our phrase enhanced indices can serve queries with fewer index operations, which greatly improves the retrieval performance.

Among the three phrase enhanced indices, the query index improves least, which may be caused by the dynamic and informality of user queries. Some queries occur rarely in the web documents, and there are also some typos and abbreviations. The hybrid index is lightly better than the document index for it makes use of the knowledge of user queries.

Our approach can also detect many site-specific phrases, newly created words, and names of people and locations. Some examples are shown in Figure 5. These out-of-vocabulary words and phrases are difficult to deal with a general, static dictionary. With the user query log, our method can find the newly rising phrases. Also the phrases from queries will be removed from the dictionary when they are out of popularity, say, after one or two months, if these phrases are never or seldom queried, they are removed from the dictionary. Thus our method can keep track of phrases popularity to make the dictionary fresh and small.

非典(SARS)
三个代表(Three Representatives)
中国青年政治学院(China Youth University for Political Sciences)
专升本(Upgrade from Junior College Student to University Student)

Fig. 5. Some Special Phrases Detected by Our Approach

5 Related Work

Much work has been done on statistics-based English phrase extraction. Shimohata et al. extract phrases from plain text according to the co-occurrence frequency of words[7]; with maximum entropy, Feng and Croft use Viterbi algorithm based on a trained Markov model to extract noun phrases dynamically[8].

For Chinese documents, the previous work includes words segmentation, dictionary construction and index terms selection. Khoo et.al. use some statistical formulas to identify multiple characters as word and evaluate many factors[10]; Yang et al. apply MI to detect word boundaries and evaluate its effect to indexing[5]. Takeda et.al. use document frequency to decide index strings[1]. Jin and Wong propose an approach to automatically construct a Chinese dictionary for IR[2]. In their approach, both local and global statistical information is considered, where local statistical information is used to select candidates and global statistical information is used to extract words based on occurrence frequency. Although this approach uses local statistical information for lower computation cost and higher accuracy, it still computes statistical information through full text just like others.

In all, there are two limitations in these previous approaches, high cost of computation and some non-meaningful extracted phrases. Different with these priori methods, our method makes use of the knowledge of user query log to select phrase candidates, which improves both precision and efficiency of retrieval.

6 Conclusion

In this paper we propose a method to extract Chinese phrases by computing MI in both user queries and web documents, which does not extract phrases through full document. It is simple to implement and the experiment results have shown its improvement on both precision and efficiency in retrieval.

Word segmentation and phrase extraction are important to character-based language IR. Using the phrases in the query log as candidates can reduce the cost of calculation and improve retrieval performance. We use these extracted phrases to construct dynamic dictionaries and build phrase enhanced indices to evaluate three MI selection methods. The result shows that the hybrid space method does the best. And our dynamic dictionaries can also alleviate the out-of-vocabulary

problem which is difficult to deal in dictionary-based natural language processing.

Although only Chinese documents are used in our experiments, we believe our approach can be easily implemented in other character-based language IR systems. We also intend to apply our approach to alphabet-based languages, such as English.

Acknowledgement. The authors are grateful to Shuguang Zhang of CERNET to provide the user query log. The authors would also like to thank the anonymous reviewers for their helpful comments in improving the paper.

References

1. Takeda, Y., Umemura, K., Yamamoto, E.: Determining indexing strings with statistical analysis. IEICE Transactions on Information and Systems **E86-D** (2003) 1781–1787
2. Jin, H., Wong, K.: A chinese dictionary construction algorithm for information retrieval. ACM Transactions on Asian Language Information Processing **1** (2002) 281–296
3. Nie, J., Briscbois, M., Ren, X.: On chinese text retrieval. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press (1996) 225–233
4. Lua, K., Gan, G.: An application of information theory in chinese word segmentation. Computer Processing of Chinese and Oriental Languages **40** (1994) 115–124
5. Yang, C.C., Luk, J.W., Yung, S.K., Yen, J.: Combination and boundary detection approaches on chinese indexing. Journal of the American Society for Information Science and Technology(JASIST) **51** (2000) 340–351
6. S.Foo, H.Li: Chinese word segmentation and its effects on information retrieval. Information Processing and Management **40** (2004) 161–190
7. Shimohata, S., Sugio, T.: Retrieving collocations by co-occurrences and word order constraints. In: Proceedings of the eighth Conference on European Chapter of the Association for Computational Linguistics. (1997) 476–481
8. Feng, F., Croft, W.: Probabilistic techniques for phrase extraction. Information Processing and Management **37** (2001) 199–200
9. Zhou, M., Tompa, F.: The suffix-signature method for searching phrase in text. Information System **23** (1997) 567–588
10. Khoo, C.S.G., Dai, Y., Loh, T.E.: Using statistical and contextual information to identify two- and three-character words in chinese text. Journal of the American Society for Information Science and Technology(JASIST) **53** (2002) 365–377