PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval

Lee-Feng Chien

Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.
Tel: 886-2-788-3799 ext. 1801
E-mail: lfchien@iis.sinica.edu.tw
Fax: 886-2-782-4814

Abstract

Considering the urgent need to promote Chinese Information Retrieval, in this paper we will raise the significance of keyword extraction using a new PAT-tree-based approach, which is efficient in automatic keyword extraction from a set of relevant Chinese documents. This approach has been successfully applied in several IR researches, such as document classification, book indexing and relevance feedback. Many Chinese language processing applications therefore step ahead from character level to word/phrase level.

I. Introduction

Under the development of global networking through the Internet, the growth in the number of electronic documents in Chinese or other oriental languages is enormous, and these have mostly being published in Japan, Korea, Singapore, Hong Kong, Mainland China, Taiwan etc. These documents are mostly non-structured and usually demand efficient IR techniques for retrieval. There is an increasing need to retrieve large numbers of such documents quickly and intelligently through world-wide information networks. Unfortunately, it is generally believed that due to the inherent differences in languages such as the lack of explicit separators, i.e. blanks or delimiters, in written oriental sentences to indicate word boundaries, the techniques developed for retrieving English documents can not be directly applied to retrieval of oriental language documents. Therefore, many researchers in oriental countries are exploring new techniques adaptable to their native languages [Wu'94, Ogawa'95, Chien'95, Nie'96, Liang'96, Lee'96]. It is the same motivation which urges researchers in Chinese regions to develop more efficient Chinese IR techniques. Considering the urgent need to promote Chinese IR, in this paper we will raise the significance of keyword extraction in Chinese IR and present a new approach, which is efficient in extracting keywords from a set of relevant Chinese documents automatically.

Automatic keyword extraction has been a critical problem in Chinese language processing. Unlike English, Chinese language does not have explicit word boundaries in written sentences. Automatic word extraction from Chinese texts is quite difficult especially for unknown words, such as names, locations, translated terms, technical terms, abbreviations etc [Chen'92]. So far, there is not many successful works on Chinese keyword extraction. However, without efficient keyword extraction, many information retrieval applications, for instance, full-text searching [Faloutsos'85], document classification[Croft'87], information filtering [Belkins'92] and text summary [Lewis'96], cannot obtain satisfactory achievements. Therefore, a new efficient keyword extraction approach which is specially useful in Chinese information retrieval applications is presented in this paper.

Traditionally, there are two types of methods to overcome problems of keyword extraction in Chinese IR. With the first kind of method, it ignores the concept of words and uses character-level information to replace word-level information in the construction of IR systems. For instance, the previous version of Csmart system [Chien'95b] is a typical example with this method. While in the second, it applies lexicon analysis to overcome this problem. For instance, there are several word-based IR systems which rely on a rigid lexicon and sophisticated word segmentation and syntactic analysis in extracting word-level information from documents [Wu'94, Nie'96]. Unfortunately, existing Chinese lexicons are often constructed for general applications and most proper nouns such as human names, which are often domain-specific keywords, are excluded from these general-domain lexicons. For instance, the team of Chinese Knowledge Information Processing (CKIP) at Academia Sinica, Taiwan, is continuing building a Chinese word lexicon with rigid syntactic information. The lexicon now contains over 100-thousand word entries and is a milestone construction for Chinese language processing, but in which only few proper nouns and domain-specific terms have been included.

In our proposed approach, we will not rely on the use of rigid lexicon and sophisticated word segmentation skills to determine either words or keywords in the text. On the contrary, an automatic statistics-based approach which is efficient in extracting significant lexical patterns (SLP) from a set of relevant documents is developed. Because most of the keywords in Chinese are not really a "word" with the

definition in linguistics without limitation of length in extracted keywords, all of the character strings in text can be theoretically candidates of keywords, which however need further analysis. In this approach, concept of significant lexical pattern is first presented, and is defined as a string consists of an arbitrary number of successive characters which are specific and significant in a certain set of relevant documents. According to this definition, for instance, for a set of documents related to Information Retrieval, a word like "布林" (Boolean), a phrase like "資訊檢索" (Information Retrieval), a longer phrase like"全文檢索技術" (Full-Text Retrieval Techniques) can be all SLP, if they are "specific" and "significant" (this will be further discussed in Section III).

The proposed approach is basically a three-step automatic process. First, in order to fast extract all of the lexical patterns without limitation of pattern length from relevant documents, an efficient working structure extended from PAT tree is created [Gonnet'92]. Using this data structure, all possible character strings with their frequency counts in a certain set of relevant documents can be retrieved and updated in a very efficient way, but not every character string with arbitrary length needs to be stored. Second, to filter out the character strings in the PAT tree which are incomplete in semantics and lack of representatives, a mutual-information-based filtering algorithm is then applied to perform detailed analysis. By means of estimating the associations of composed substrings for possible lexical patterns, this algorithm produces effective performance in filtering less significant patterns. Third, in determination of final SLP, a refined method based on a common-word lexicon, a general-domain corpus and a keyword determination strategy are also presented. Lexical patterns which are not really specific will be further removed in this step.

Based on the proposed approach, exciting results have been achieved in different applications, such as book indexing, document classification and relevant feedback in information retrieval. The obtained results show that, if a certain set of relevant documents can be given, the keywords without limitation of string length can be effectively extracted using this approach, especially for Chinese proper nouns and patterns longer than two characters. Besides, the proposed approach is very easy to implement and the reliance of rigid lexicon and sophisticated word segmentation skills can be effectively reduced. With this approach for keyword extraction, many Chinese language processing applications therefore step ahead from character level to word/phrase level, especially on information retrieval applications.

In the rest of this paper, an overview of the proposed keyword extraction approach will be first introduced in Section II. In Sections III and IV, the concept of PAT tree and the mutual-information-based extraction of SLP will be described respectively. In addition, experiments and applications on book indexing, document classification and relevance feedback will be discussed in Section V. Finally, concluding remarks will be given in Section VI.

II. Overview of the Proposed Approach

As mentioned in the previous section, the proposed approach is a three-step automatic process. For the convenience of further discussion, Fig. 1 shows an abstract diagram of the proposed approach, in which it can be found that the three steps are lexical patterns construction, initial SLP extraction and refined SLP extraction.

At first, the text collection for keyword extraction is assumed to be a certain set of relevant documents. This is because SLP supposedly will occur at least several times in the text collection. This is, however, the restriction and basic requirement of the proposed approach.

To avoid the loss of unknown proper nouns, at the step of lexical pattern construction, all of the lexical patterns with an arbitrary length at sentence level in the text collection will be recorded. In this way, for a text collection with a phrase "關鍵詞抽取" (keyword extraction) for instance, the composed character substrings like "關"、"關鍵"、"關鍵詞"、"關鍵詞抽"、"關鍵詞抽取"、"鍵"、"鍵詞"、"鍵詞抽"、"鍵詞抽取"、"詞"... etc. are all possible candidates of SLP and need to be efficiently recorded (although only two phrases "關鍵詞" and "關鍵詞抽取" are supposed to have complete word boundaries and semantic meanings). For recording the occurrences of each identical lexical patterns without the limitation of length and extracting statistical information such as co-ocurrence frequencies of lexical patterns in the text, the PAT tree data structure is adopted and extended. The PAT tree used is economic in reducing space overhead and efficient in storing and retrieving character string.

In addition, to be able to extract SLP with correct and complete word boundary or segmentation such as "關鍵詞索引"、"自然語言"、"中央研究院"、"文字資料庫", rather than"關鍵詞索"、"然語言"、"中央研究"、"字資料庫", at the

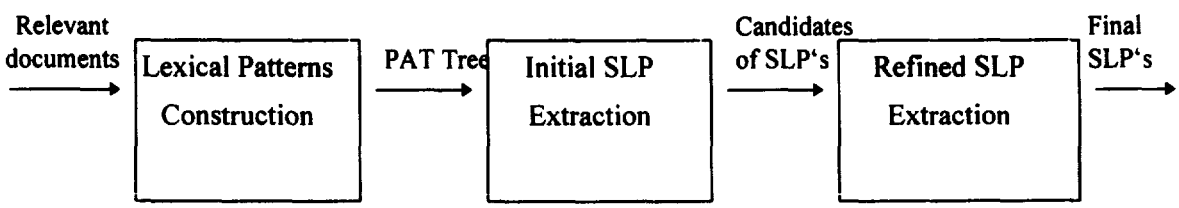

Fig.1 An abstract diagram to show the process of the proposed keyword extraction approach

step of initial SLP extraction, an effective SLP estimation method and filtering algorithm are developed. By means of estimating the mutual information among composed substrings for each possible lexical patterns, most of the lexical patterns which are incomplete in semantics and lack of representatives in the PAT tree can be filtered out (for example, patterns such as "關"、"關鍵"、"關鍵詞抽"、"鍵"、"鍵詞"、"鍵詞抽"、"鍵詞抽取"、"詞"... etc extracted from the phrase "關鍵詞抽取" can be filtered out). Furthermore, at the step of refined SLP extraction, the remaining lexical patterns will be further examined using a common-word lexicon, a general-domain corpus and a keyword determination strategy. Only lexical patterns which are really specific and significant will be extracted as the final SLP.

III. The PAT Tree

PAT tree is an efficient data structure successfully used in the area of information retrieval. It was developed by Gonnet [Gonnet'92] from Morrison's PATRICIA algorithm (Practical Algorithm to Retrieve Information Coded in Alphanumeric) [Morrison'68] for indexing a continuous data stream and locating every possible position of a prefix in the stream. The PAT tree is conceptually equivalent to compressed digital search tree but smaller. The superior features of the PAT tree data structure is most resulted from the use of so-called semi-infinite strings [Manber'91] in storing the substream values in the nodes of the PAT tree. Using this data structure to index the full-text of documents, all possible character strings, including their frequency counts in the documents, can be retrieved and updated in a very efficient way, yet not every character string with arbitrary length is needed to be stored.

When applying PAT tree to Chinese keyword extraction, instead of recording the semi-infinite strings at document level, we record them at sentence level. In this way, we use punctuation marks such as "." and "," as delimiters to determine a sentence boundary. For example, the semi-infinite strings generated for the string "詞彙自動抽取，減化索引困難" at the sentence level are:

詞彙自動抽取000...
彙自動抽取00000...
自動抽取0000000...
動抽取000000000...
抽取00000000000...
取0000000000000...
減化索引困難000...
化索引困難00000...
索引困難0000000...
引困難000000000...
困難00000000000...
難0000000000000...

The adopted PAT tree is basically a data construction for the composed semi-infinite character strings of a given data stream at the sentence level. These semi-infinite strings logically have the same length (the length is dependent on the largest size of sentences in the corpus and the short strings are inserted "0" bits in the end). For each data stream recorded in the PAT tree, the existence of its composed substrings can be easily detected if the searching string is exactly a prefix of a semi-infinite string. For example, to check if "動抽" and "彙自動抽 " are substrings of data stream "詞彙自動抽取" with the PAT tree, the answer is yes because "動抽" can be found as a prefix of the semi-infinite string "動抽取000000000" and "彙自動抽 " that of the semi-infinite string "彙自動抽取00000".

During implementation, each identical semi-infinite string is represented as a node in the PAT tree and has a pointer to the position in document to save space. For example, for the data stream "個人電腦, 人腦" shown in Fig. 2, nodes 0, 2, 4, 6, 9 in the PAT tree exactly represent the occurrence and the data positions of the identical semi-infinite strings "個人電腦", "人電腦", "電腦", "腦" and "人腦" in the data stream respectively. In the mean time, each node consists of three parameters: comparison bit, # of external nodes and frequency count for the purpose of searching and information update. The # of external nodes indicates the number of external nodes in the subtrees and the frequency count indicates the frequency of the corresponding character string occurred in the data stream. Moreover, the comparison bit indicates the first different bit of the character strings recorded in the subtrees. The comparison bit is primarily used in each node as an indication of which bit of the searching string is to be used for branching. A zero bit will cause a branch to the left subtree, and a one bit will cause a branch to the right subtree. By storing the frequency of recorded character string and the number of external nodes (those nodes who have comparison bits greater than their parents') of each node, we are able to know the frequencies of every strings stored in the PAT tree, and this makes it possible to extract all of the keywords automatically. For instance, to search for "電腦" in the example PAT tree (its Chinese codes is 10111001 01110001 00000000 ... and the bits are numbered from 1, 2, 3, as that shown in Fig. 2), it first checks node 0 by comparison bit 0 and goes to node 4 by default zero bit. Then, it checks comparison bit 4 and goes to node 6 by obtaining one bit. Finally, it checks comparison bit 8 and goes to node 4 again. It matches the searching string with the corresponding character string of node 4 and stops the search.

Basically, although PAT tree is a very efficient data structure to record all of the substrings of documents, it really demands large space overhead and takes time to build. In Section V, some empirical results on the construction of PAT tree will be reported.

IV. Extraction of Significant Lexical Patterns

Estimation of Significant Lexical Patterns

As described in Section II, further analysis is necessary to filter out the lexical patterns in the PAT tree, for example phrases,

"關鍵詞抽" and "鍵詞抽取", which are incomplete in semantics and lack of representatives. According to the findings in our experiments, most of the significant lexical patterns have strong association between its composed and overlapped substrings. Significant patterns extraction is, therefore, done by observing mutual information of two overlapped patterns with the following significance estimation function SE:

$$SE_c = MI_{ab} = \frac{\Pr(c)}{\Pr(a) + \Pr(b) - \Pr(c)}$$

$$= \frac{\frac{f_c}{F}}{\frac{f_a}{F} + \frac{f_b}{F} - \frac{f_c}{F}}$$

$$= \frac{f_c}{f_a + f_b - f_c} \qquad (1)$$

where c is the lexical pattern to be estimated, $c = c_1, c_2, ..., c_n$, a and b are the two longest composed substrings of c with the length n-1 , i.e., $a = c_1, ..., c_{n-1}$, $b = c_2, ..., c_n$, f_a , f_b and f_c are the frequencies of a , b and c in the PAT tree, respectively, and F is the total frequency. Conventionally, mutual-information-based estimation is used in natural language processing for estimating the possibility of a pair of words to be formed as a new phrase. Since the main purpose in our estimation is to judge if pattern c is more complete in semantics than its composed substrings (especially on longest composed substrings), different from conventional estimation the examining string pair a and b are defined to be the two longest composed and overlapped substrings of c. Although this kind of estimation looks simple and straightforward, it is really useful in filtering incomplete character strings and efficient in reduction of computational complexity. Basically, if SE_c is large, it can be found that most of the time patterns a and b have to occur together in the text collection. It seems to indicate that, pattern c is more complete in semantics than either a or b. This can be further clarified with the following illustration:

a = "關鍵词抽" f_a =6

b = "鍵词抽取" f_b =6

c = "關鍵词抽取" fc =6

$SE_c = MI_{ab}$=1

This example shows that "關鍵词抽"and "鍵词抽取" (incomplete lexical patterns) always come together (MI_{ab}=1), consequently, "關鍵词抽"and "鍵词抽取" seem to be incomplete in semantics than "關鍵词抽取". Thus, both incomplete phrases can be filtered out. However, in some cases it will not work simply based on the above mutual-information-based estimation. For instance, some significant patterns such as "關鍵词" (keyword) might not have a large SE value. That is because the composed substring pattern "關鍵" (key) is a frequently-used word. In this case, a refined filtering algorithm based on the above SE measurement and the PAT-tree structure is further presented.

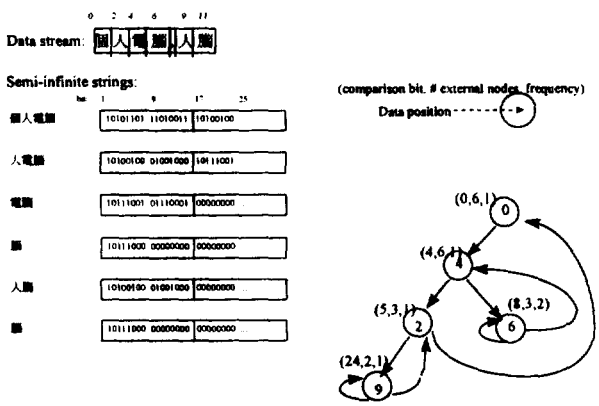

Fig 2. Example of the Chinese PAT tree

The PAT-Tree-Based Filtering Algorithm

The procedure of the filtering algorithm is given below and a corresponding abstract diagram shown in Fig. 3:

For each lexical pattern a in the PAT tree (like node A in Fig. 3, where node A indicates the set of character strings with the prefix "關鍵词抽"), where a has not been marked as a non-SLP and $f_a >= TH_f$ (TH_f is a pre-defined constant of frequency value to avoid too much computation), there are three cases to check if a is a candidate of SLP by considering one of its successor c (like node C which indicates the set of character strings with the prefix "關鍵词抽") and searching for the other overlapped subpattern b (like node B which indicates the set of character strings with the prefix "鍵词抽取"):

(I) if a is a terminal node , then a is determined as a candidate of SLP (for

example, a= "關鍵詞抽取")

(II) if $SE_c >= TH_{se}$ (TH_{se} is a pre-defined threshold), then *a* and *b* are determined non-candidates of SLP and will have a non-SLP mark on the PAT-tree. In addition, if $f_c >= TH_f$, the above process will be processed recursively from *c* to check if *c* is a candidate of SLP. (for example, a = "關鍵詞抽", b ="鍵詞抽取", c= "關鍵詞抽取")

(III) if $SE_c < TH_{se}$, but $f_a >> f_b$ then *a* is determined as a candidate of SLP, while *b* and *c* are uncertain. In addition, if $f_c >= TH_f$, the above process will be processed recursively from *c* to check if *c* is a candidate of SLP. (for example, a = "關鍵詞", b ="鍵詞抽", c= "關鍵詞抽")

Basically, the above is a recursive process which will be started from the longest patterns with frequency larger than the threshold and then backtraced to their ancestors. Since the definition of SE function has carefully considered the characteristics of the PAT tree, the above process is very easy to be implemented using the PAT tree structure. For instance, it is not difficult to traverse each node in the PAT tree. So as to, every node (every identical lexical pattern) can be estimated. At the same time, all of the necessary frequency parameters such as f_a, f_b and f_c are easy to be extracted on the PAT tree, as described in the last section.

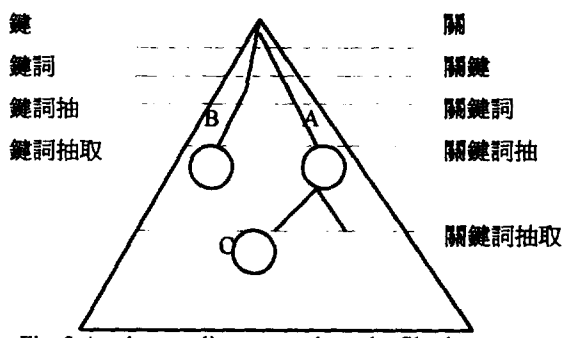

Fig. 3 An abstract diagram to show the filtering process of SLP extraction on the PAT tree.

Filtering of Incomplete Lexical Patterns

In fact, the lexical patterns extracted with the above algorithm depend mainly on the context rather than the frequencies of the patterns. Therefore, a lexical pattern is extracted as long as it appears in the content and its *SE* value is large enough. At the same time, the above filtering algorithm is effectively to remove most of incomplete lexical patterns and extract the candidates of SLP. For example, the incomplete lexical patterns, such as "關鍵詞抽" and "鍵詞抽取" can be removed, but "關鍵詞" and "關鍵詞抽取" can be extracted. In our findings of the experiments, the above filtering algorithm is especially useful in extracting SLP like names, locations and technical terms. This is because most of Chinese names are frequently three-character phrases such as "李登輝" (Den-Huei Li) and technical terms are four-character phrases such as "資訊檢索" (Information Retrieval), the obtained SE values are not small (for instance, both MI"[illegible]", "[illegible]" and MI"[illegible]", "[illegible]" are supposedly equal to 1).

The Refined Procedure

However, there do exist weaknesses with the above filtering algorithm. Firstly, even a candidate of SLP has a large SE value, it cannot show its significance in the text collection. Secondly, many candidates of SLP are commonly found in our daily use. Hence, all of the candidates of SLP extracted initially should be further filtered using the following refined procedure.

With the refined procedure, all of the candidates of SLP will be checked using a common-word lexicon and a Chinese PAT tree constructed by using a general-domain corpus. If a candidate of SLP appears either in the common-word lexicon or in the candidate list of SLP of the general-domain PAT tree, it is treated as a non-specific candidate and will be removed from the list of final SLP. The remaining candidates of SLP will be further checked by observing their frequencies and distributions in the text collection. Only significant patterns will be extracted as final SLP.

V. Experiments and Applications

Book Indexing

Basically, it is not easy to evaluate the performance of any keyword extraction approaches. In order to realize the performance of the proposed approach, an application of book indexing was performed. The experiments on this application were done to extract keywords from a book in the area of information retrieval, which contains 1 MB text file and consists of a total about 200,000 Chinese characters. The book had been indexed manually first and a total of 190 keywords extracted. To compare with the performance of manual and automatic keyword extraction, some of the test results obtained in this application are shown in Table 4. Since some keywords extracted are similar to manual keywords (with the same meaning but having one or two characters mismatched), both of the exact match and near match methods are used in the performance estmation. It has to point out that there have no satisfactory results in extracting Chinese book indexing so far. So, from this table, it can be found that the obtained recall and precision values look promise of effectiveness. In addition, according to our analysis, though 70% of extracted keywords didn't exactly appear in the set of manual keywords, many of them are actually domain-specific terms. These keywords are still important to be extracted in many IR applications. Besides, the extracted keywords almost possess correct word boundary. This indicates that the proposed approach works well in elimination of a significant number of incomplete lexical patterns. In addition, it is worthy of notice that most of the keywords missed are due to their low frequency values in this book. This reveals an important direction for further improvement.

Document Classification

Since the above experiments show that most of the extracted keywords are domain-specific terms (even though they are not author-defined keywords), it is believed that the proposed

approach would be also useful in document classification. Automatic document classification is generally done through processing textual data with discriminating analysis on keywords of documents. There are two phases to classify documents: first, defining the keyword set, and second, analyzing the documents. Because words in lexicon are too common to be "key" words, it is difficult to define a proper keyword set. With the PAT tree data structure and the proposed keyword extraction approach which may generate keywords dynamically from a set of relevant documents, we are able to define a more proper keyword set than that defined by a static lexicon.

Character Size of Keywords	2	3	4	5	6	>6	Avg.
Number of Manual Keywords (Correct Keywords)	5	10	44	27	46	58	
Number of Extracted Keywords	4	26	106	50	62	28	
Number of Correct Keywords Extracted (Exact Match)	2	4	22	14	21	19	
Obtained Precision (Exact Match)	0.5	0.15	0.21	0.28	0.33	0.68	0.30
Obtained Recall (Exact Match)	0.4	0.4	0.5	0.51	0.45	0.33	0.43
Number of Correct Keywords Extracted (Near Match)	2	4	30	19	29	22	
Obtained Precision (Near Match)	0.5	0.15	0.28	0.38	0.48	0.79	0.38
Obtained Recall (Near Match)	0.4	0.4	0.68	0.70	0.63	0.38	0.56

Table 4. Some of the test results obtained in Chinese book indexing application

The results of a preliminary experiment on Chinese document classification based on the proposed approach are shown from Table 5 through Table 7. The experiment uses daily news articles from Central News Agency in Taiwan which are posted on the Internet and grouped manually into eight sections, namely section 1: congress/politics, section 2: judiciary/transport, section 3: economics/finance, section 4: education, section 5: recreation, section 6: local government, section 7: weather report, section 8: misc. In the initial step, we first trained eight sets of keywords corresponding to the eight groups. The training corpus were obtained from news articles published on August 1-16, 1996. The keywords for each section were extracted with the approach proposed in this paper. Then we also extracted keywords from test corpus, which were news articles of the whole eight sections published on August 17, 19 and 20. It has to note that this experiment was performed to realize the effectiveness of the proposed keyword extraction approach rather than to develop a really efficient method in Chinese document classification at that stage. The discriminating analysis is, therefore, completed with simple set operation as shown in (2).

$$SI_{i,j} = \frac{\text{card}(S_i \cap A_j)}{\text{card}(A_j)} \tag{2}$$

Where S_i is the keyword set of training articles in section i, A_j is the keyword set of test articles in section j. *SI* defines the similarity between test articles and training articles in different sections.

Similarity	Section 1	Section 2	Section 3	Section 4	Section 5	Section 6	Section 7	Section 8
A_1	**0.4242**	0.3030	0.2727	0.2424	0.1515	0.2121	0	0.3333
A_2	0.1896	**0.4482**	0.2413	0.2068	0.1206	0.1896	0.0172	0.2586
A_3	0.3076	0.4615	**0.5384**	0.3076	0.3846	0.2307	0.0769	0.1538
A_4	0.0857	0.1285	0.0857	**0.2428**	0.1142	0.0857	0.0142	0.1285
A_5	0.1818	0.2727	0.2727	0.1818	**0.6363**	0.2727	0	0.2727
A_6	0.3076	0.3846	0.1538	0.3076	0.2307	**0.4615**	0.0769	0.1538
A_7	0.0217	0.0434	0	0	0	0	**0.5869**	0.0217
A_8	0.1666	0.0833	0.0833	0	0	0	0.0833	**0.4166**

Table 5. Similarities for news on Aug. 17

Similarity	Section 1	Section 2	Section 3	Section 4	Section 5	Section 6	Section 7	Section 8
A_1	**0.6219**	0.2195	0.2195	0.2560	0.1219	0.2195	0.0121	0.3780
A_2	0.1428	**0.5714**	0.2142	0.2142	0.2857	0.2857	0	0.2142
A_3	0.2666	0.2	**0.4666**	0.1333	0.2	0.2	0	0.0666
A_4	0.1363	0.1818	0.1363	**0.2727**	0.1818	0.1363	0	0.0454
A_5	0.0638	0.0851	0.0638	0.0425	**0.1702**	0.1063	0.0212	0.0851

A_6	0.1754	0.2280	0.1052	0.1052	0.0526	**0.4561**	0	0.1403
A_7	0	0.0303	0	0	0	0	**0.5757**	0.0303
A_8	0.2291	0.1562	0.1354	0.1145	0.0937	0.1458	0.0208	**0.4166**

Table 6. Similarities for news on Aug. 19

Similarity	Section 1	Section 2	Section 3	Section 4	Section 5	Section 6	Section 7	Section 8
A_1	**0.5730**	0.2808	0.2584	0.2134	0.1573	0.1797	0	0.3820
A_2	0.2068	**0.4827**	0.3103	0.1724	0.1724	0.2413	0.0344	0.1724
A_3	0.2317	0.1219	**0.4512**	0.1219	0.0853	0.0853	0.0243	0.1707
A_4	0.0540	0.0945	0.0810	**0.1351**	0.1081	0.0810	0	0.0540
A_5	0.1315	0.1578	0.1578	0.1052	**0.3421**	0.0526	0	0.1578
A_6	0.1562	0.25	0.1406	0.125	0.0781	**0.3281**	0.0156	0.125
A_7	0.0208	0.0416	0	0	0	0	**0.4791**	0.0208
A_8	0.2631	0.1789	0.2105	0.1894	0.1684	0.1263	0.0315	**0.4421**

Table 7. Similarities for news on Aug. 20

The above three tables show that it is promising to achieve automatic document classification with the proposed keyword extraction approach using a simple similarity function. It is also observed that the more precise the pre-divided groups, the better the results can be. The weather section (section 7) is an example with a very high precision. Actually, more sophisticated algorithms can be developed from this point. For example, eliminating common words which belong to all or most of the groups or applying other discriminating analysis functions, such as perplexity functions. Besides, in this application, it also shows the processing speed and space requirement with the proposed approach. For instance, Table 8 shows some of the space and speed performance runned on SUN SPARC 20 in the training phase.

Section	Text size (KB)	PAT tree size (KB)	Time of building PAT tree	Time of extracting keywords	Number of extracted keywords
1	1,682	11,657	2:34'43"	13'35"	3,165
2	1,736	12,300	2:47'46"	9'18"	2,665
3	1,276	7,435	1:37'10"	5'36"	2,043
4	517	3,873	49'35"	2'32"	632
5	358	2,487	37'10"	1'43"	704
6	1,206	8,654	1:56'23"	6'44"	1,797
7	1,698	615	26'03"	38"	792
8	1,479	4,383	51'38"	4'04"	1,721

Table 8. Some of the speed and space performance obtained in the training phase for document classification application

Relevance Feedback

Keyword extraction technique can be also applied to implementing relevance feedback function on information retrieval systems to cpature users' minds. Putting the system with the extracted keywords as expanded queries instead of the whole documents may get retrieved results closer to users' needs. In order to pursue high performance Chinese information access on the Internet, an IR system called Csmart has bee developed at Academia Sinica to apply natural language information retrieval techniques to Chinese networked information discovery and retrieval [Chien'95b, Chien'97]. The design of the IR system is completely based on a two-stage searching concept and character-level signature file approach, whihc is fast and intelligent in the retrieval of large Chinese full-text document databases. Using this approach, the inherent difficulties of Chinese word segmentation and proper noun identification have been effectively reduced, and efficient quasi-natural language queries and relevance feedback functions implemented. For realizing the performance of using the proposed keyword extraction approach in such an application of IR systems, the Csmart system has been extended and tested. Fig. 4 shows this extension, where it consists of a keyword extraction phase in the construction of relevance feedback function (it also needs to point out that since there is no effective keyword extraction approach, as we know, no existing Chinese IR systems provide such a relevance feedback function). On a test of 100 queries in searching for a news database with about 100,000 pieces of news. We estmated the obtained recall and precision values by using the initial quasi-natural language queries, Csmart's original relevance feedback function, and the improved relevance feedback function. The test

results show that better performance can be achieved with the proposed approach in implementing relevance feedback function. Right now, the improved function has been opened with the Csmart system to provide Chinese IR service on the Internet (http:// csmart2.iis.sinica.edu.tw).

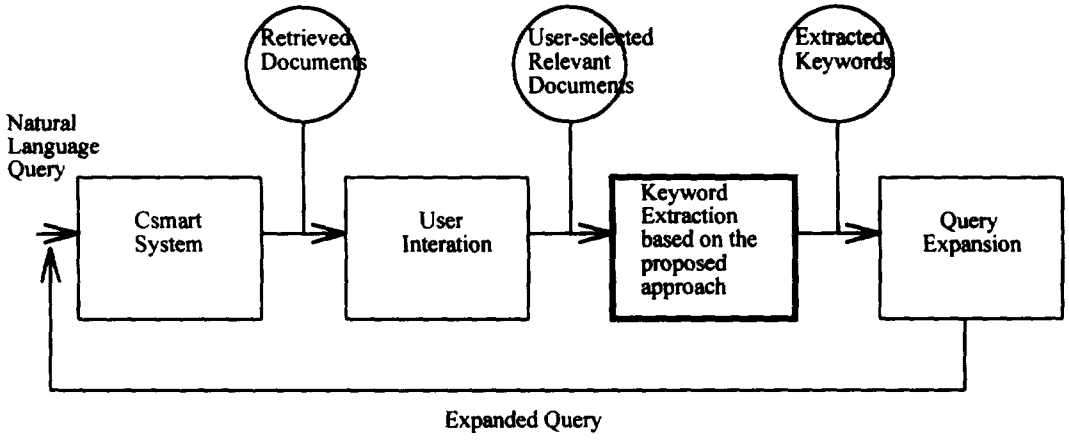

Fig. 4 An abstract diagram to show the process of the improved relevance feedback fucntion with the Csmart system.

VI. Conclusion

The proposed Chinese PAT tree does reduce the difficulty of keyword extraction in Chinese, which is critical and fundamental. Using this data structure all possible character strings can be easily retrieved and updated, and the mutual-information-based filtering algorithm can be performed. Keywords, in special proper nouns which were excluded in the general lexicon, are therefore possible to be extracted. Besides, lexical patterns which are incomplete in semantics and lack of representatives can be effectively filtered out. Moreover, the proposed approach is very easy to implement and the reliance on rigid lexicon and sophisticated word segmentation skills can be reduced.

Though many researches are blocked by the problem of Chinese keyword extraction, now these may have a new solution. Better performance in document classification, book indexing and relevance feedback have been achieved in our research. We believe that there are many other applications including language parsing, natural language processing, keywords indexing, and so on, which are suited to adopt the proposed keyword extraction approach. However, more in-depth discovery of the Chinese PAT tree and the proposed approach need to be further investigated.

References

1. [Chang'92] Jyun-Sheng Chang, Tsung-Yih Tsengm Ying Chen, Huey-Chyun Chen, Shun-Der Chen, John S. Liu and Sur-Jin Ker, "A Corpus-based Statistical Approach to Automatic Book Indexing", Proceedings of the 3rd Conference on Applied Natural Language Processing (ANLP'92), pp. 147-151, Trento, Italy, 1992.

2. [Belkin'92] Nicholas J. Belkin, et al., "Information Filtering and Information Retrieval: Two Sides of the Same Coin ?", Communications of the ACM, Vol. 35, No. 12, Dec. 1992.

3. [Chen'92] Keh-Jiann Chen et al., "Word Identification for Mandarin Chinese Sentences", COLING'92.

4.[Chien'95a] Lee-Feng Chien, 'Fast and Quasi-Natural Language Search for Gigabytes of Chinese Texts", ACM SIGIR' 95.

5.[Chien'95b] Lee-Feng Chien, "尋易(Csmart) -- A High-Performance Chinese Document Retrieval System", Proceedings of the 1995 International Conference on Computer Processing of Oriental Languages, ICCPOL'95, pp. 176-183.

6. [Chien'97] Lee-Feng Chien et al., "Networked Chinese Information Access Using Speech and Natural Language Information Retrieval Techniques", Proceedings of the 1997 International Conference on Computer Processing of Oriental Languages, ICCPOL'97, pp. 669-674.

7. [Croft'87] W. B. Croft, "Clustering Large Files of Documents Using the Single-Link Method:, JASIS, 35, pp. 268-76.

8. [Faloutsos'85] C. Faloustos, "Access Methods of Text", ACM Computing Surveys, 17(1), pp. 49-74.

9.[Gonnet 92] Gaston H. Gonnet, Ricardo A. Baeza-yates and Tim Snider, "New Indices for Text: Pat Trees and Pat Arrays", Information Retrieval Data Structures & Algorithms, Prentice Hall, pp. 66-82, 1992.

10.[Lee'96] Lee, Ahn and Shin, "An Effective Indexing Method for Korean Text Retrieval", International Workshop on Information Retrieval with Oriental Languages, Korea, 1996.

11. [Lewis'96] David D. Lewis and Karen Sparck Jones, "Natural Language Processing for Information Retrieval", Communications of the ACM, Vol. 39, No. 1, Jan. 1996, pp. 92-101.

12.[Liang'96] Tyne Liang, Suh-yin Lee and Wei-Pang Yang, "Optimal Weight Assignment for a Chinese Signature File", Information Processing and Management, Vol 32, No. 2, pp. 227-237, 1996.

13.[Manber 91] Manber, U. and R. Baeza-Yates, "An Algorithm for String Matching with a Sequence of Don't Cares", Information Processing Letters, 37, pp.133-136, 1991.

14. [Morrison 68] Morrison, D., "PATRICIA :Practical Algorithm to Retrieve Information Coded in Alphanumeric", JACM, pp. 514-534, 1968.

15. [Nie'96] Jian-Yun Nie et al., "On Chinese Text Retrieval:", ACM SIGIR'96.

16. [Ogawa'95] Y. Ogawa, "A New Character-based Indexing Organization Using Frequency Data for Japanese Documents", ACM SIGIR'95.

17. [Wu'94] Zimin Wu and Gwyneth Tseng, "Chinese Text Segmentation for Text Retrieval:Achievements and Problems". JASIS, 44(9), 1994, 532-542

SIGIR 97 Philadelphia PA, USA