Chinese Keyword Extraction Based on Max-duplicated Strings of the Documents

Wenfeng Yang
Electronic Engineering Department
Tsinghua University, P.R.China
yangwenfeng97@mails.tsinghua.edu.cn

Xing Li
Electronic Engineering Department
Tsinghua University, P.R.China
xing@cernet.edu.cn

ABSTRACT
The corpus analysis methods in Chinese keyword extraction look on the corpus as a single sample of language stochastic process. But the distributions of keywords in the whole corpus and in each document are very different from each other. The extraction based on global statistical information only can get significant keywords in the whole corpus. Max-duplicated strings contain the local significant keywords in each document. In this paper, we designed an efficient algorithm to extract the max-duplicated strings by building PAT-tree for the document, so that the keywords can be picked out from the max-duplicated strings by their SIG values in the corpus.

Categories & Subject Descriptors: H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing – Dictionaries

General Terms: Algorithms

Keywords: MDS, Chinese keyword extraction, PAT tree

1. INTRODUCTION
With the fast development of the Internet, the number of electronic documents in Chinese becomes enormous. There are increasing needs to retrieve, filter and mine Chinese text through world-wide information networks. Unfortunately, Chinese text is different from English text in that there is no explicit word boundary. The techniques developed for English text processing can not be directly applied to Chinese text processing. Chinese text is made up of ideographic characters, and a word can comprise one, two or more such characters, without explicit indication where one word ends and another begins. The words out of the vocabulary, such as names, locations and translated terms, are difficult to be detected in Chinese text. Automatic keyword extraction is a critical problem in Chinese language processing.

Sproat and Shih [1] developed a purely statistical keyword extraction method that utilizes the mutual information between two characters in the corpus: $I(x,y) = \log p(x,y) - \log p(x) - \log p(y)$. The character pairs with the largest mutual information are assumed to be keywords. The limitation of this method is that it can only deal with keywords of 2 statistical units. In order to extract the keywords of arbitrary length, Chien [2] designed a PAT-tree-based method. The PAT tree can provide indices to all possible streams of characters in the corpus. All of the required statistical parameter can be extracted directly from the PAT tree. All the keyword extraction methods based on the global statistical analysis of the whole collection are easy to be impacted by the frequency fluctuation of keywords with the same substrings. For example, the famous name "曹雪芹" occurs 350 times in our 600M corpus, the total frequency of the other name beginning with "曹雪" is only 41 times. The global statistical method can not efficiently extract these unsignificant keywords, such as "曹雪华", "曹雪涛". In this paper, we present a keyword extraction method based on string frequency analysis in documents. Firstly, we designed a modified PAT-tree building algorithm to get the local significant strings in each document, then use the distributions of these strings in the corpus to select the eligible keywords.

SIGIR'02, August 11-15, 2002, Tampere, Finland.
ACM 1-58113-561-0/02/0008.

2. KEYWORD EXTRACTION
2.1 Word Segmentation
The training corpus used in our experiments consisted of news articles from People Daily and some other electronic novels. The data-set consisted of approx. 277,000 documents. Word segmentation is a necessary step to bring language knowledge into statistical approaches and can make the max-duplicated strings more meaningful. Forward and backward maximal matching algorithms are used in word segmentation. On the purpose of keyword extraction, there are no needs to deal with the ambiguities of word segmentation. The single characters in the ambiguous parts of the documents are treated as one-character words.

2.2 Max-duplicated String
Let S stand for the string whose beginning and ending are word boundaries in a segmented document, $f(S)$ is the frequency of the string S, $|S|$ is the length of string S, and $S \subset S'$ means that the string S' has a sub-string S. Then max-duplicated string is the string S, whose $f(S)>1$ and the frequencies $f(S')$ of all the strings S' containing the string S is smaller than $f(S)$.

2.3 Modified PAT-tree Building Algorithm
PAT tree is an efficient data structure successfully used in information retrieval. It was developed by Gonnet[3] from Morrison's PATRICIA algorithm. Since Chinese documents are segmented by words and punctuation marks, the semi-infinite strings can be generated at word level. For example, The string "自然语言处理" are segmented by the words "自然", "语言" and "处理". The semi-infinite strings generated for the string at word level are:

"自然语言处理 000…"
"语言处理 0000000…"
"处理 00000000000…"

A semi-infinite string is recorded by one path from the root to the leaf in the PAT tree. All the semi-infinite strings with the same prefix have the same path from the root to an internal node. In the process of building a PAT tree for one document, all the strings with $f(S) > 1$ can be detected. In order to keep the suffix relation of the semi-infinite strings for one sentence, the semi-infinite strings must be inserted to the PAT tree by their original orders in the sentence. If one duplicated string S_1 is found by inserting one semi-infinite string, the next duplicated string S_2 found by next semi-infinite string should be carefully checked. There are three cases:

(1) $S_2 \subset S_1$ and $f(S_2)=f(S_1)$, then S_2 is not a max-duplicated string.
(2) $S_2 \subset S_1$ and $f(S_2)>f(S_1)$, then S_2 is a max-duplicated string.
(3) $S_2 \not\subset S_1$, then S_2 is a max-duplicated string.

2.4 Filtering Max-duplicated Strings

Noises are unavoidable in the keyword extraction. Some max-duplicated strings are produced by high frequency keywords randomly attaching function words, such as "的", "以" and "用". The frequencies of those max-duplicated strings are much lower than its central keywords'. We define the stability of a max-duplicated string as *stab(S)*:

$$stab(S) = f(S) * |S|$$

Similar to word segmentation, we then segment the document once again by these max-duplicated strings. In the ambiguous segmentations of the document, only the max-duplicated string with the largest stability can be chosen as a suitable segmentation. After the document is segmented, those max-duplicated strings not being selected in any segmentation should be filtered out.

2.5 SIG Parameter and Selecting Keywords

The distributions of max-duplicated strings in the corpus represent its specialization. If one max-duplicated string occurs in several documents with high frequency, then the max-duplicated string has high specialization and is a good keyword candidate. By our hypothesis, the typical distributions of keyword and extraction noise are shown in Figure 1. Because the extraction noise has a high document frequency and its frequency in single document is low, its distribution curve is smooth. On the contrary, the keyword has a sharper distribution curve.

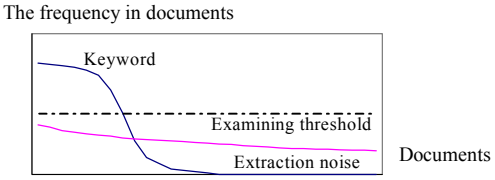

Figure 1: The distribution curves of keyword and noise

TOTAL(S) is the total frequency of the string S in the corpus and stands for the whole area under the distribution curve. If the value of the examining threshold is n, then the $FREQ_n(S)$ is the total frequency of the string S which occurs at least n times in one document and stands for the area under the distribution curve but above the threshold line. The significance of the string in corpus can be evaluated by the *SIG* parameter:

$$SIG_n(S) = FREQ_n(S) / TOTAL(S)$$

Since the max-duplicated string at least occurs 2 times in documents, the default of examining threshold is 2. After the SIG values of all the max-duplicated strings calculated, we can sort these max-duplicated strings by their SIG value. The strings with high SIG values are good keywords and those with low SIG values are assumed as extraction noises.

3. RESULTS

The size of the data-set used in our experiment is about 600M. In order to examine the effects of the word segmentation, we firstly tried keyword extraction without word segmentation and max-duplicated string filtering. We got 20,449,484 max-duplicated strings. The precision of keyword extraction is about 50.7%. After having segmented the documents by one 50,000-word dictionary, we got 17,247,407 max-duplicated strings and the precision is improved to 59.3%. The language knowledge in the dictionary is very helpful to keyword extraction.

Secondly, we filtered the 17,247,407 max-duplicated strings using the max-duplicated string filtering algorithm. We got 6,233,217 stable max-duplicated strings. The precision is up to 76.4%. Max-duplicated string filtering can greatly improve the precision of the extraction.

In the 6,233,217 max-duplicated strings there are about 430,000 unique strings. After removing the strings whose frequencies are less than 30, we got 200,199 keyword candidates and 30,153 of them are in the dictionary. We sorted the 170,046 candidates out of the dictionary by their SIG value from up to down and get a sorted candidate list. The precision of keyword extraction is proportional to candidates' SIG values (Figure 2). The candidates with the largest SIG value have the precision near to 100%. But at the end of the list, the candidates with small SIG values are almost extraction noises. Using the SIG value, we can easily select the good keywords from these candidates.

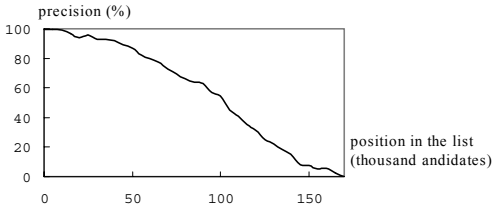

Figure 2: The precision curve of the sorted candidate list

4. CONCLUSION

We present a statistical approach to extract the keywords on base of the max-duplicated strings of the documents in the corpus. The results show that the word segmentation and max-duplicated string filtering algorithm are very helpful to the extraction and the SIG parameter provides a way to control the quality of the keywords.

5. REFERENCES

[1] R.Sproat, C.Shih. A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese and Oritental languages, Vol.4, 336-351, 1990

[2] Lee-Feng Chien, PAT-tree-based Keyword Extraction for Chinese Information Retrieval, ACM SIGIR'97, 50-59.

[3] Gaston H.Gonnet, Ricardo A.Baeza-yates and Tim Sinder, New Indices for Text: Pat Tree and Pat Arrays, Information Retrieval Data Structure & algorithm, Prentice Hall, pp.66-82, 1992