AUTOMATIC KEYWORD EXTRACTION OF CHINESE HOMEPAGE FOR WEB UNDERSTANDING

*WANG Jhing-Fa and **HUANG Chieh-Yi
*Department of Electrical Engineering
**Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan, R.O.C.
Tel.+886-6-2746867, FAX:+886-6-2746867,
E-mail: wangjf@server2.iie.ncku.edu.tw, chiehyi@ms12.hinet.net

ABSTRACT

This paper proposes a method to extract keyword from Web pages automatically. The automatic keyword extraction technique is a very important part for the document and web understanding on the Internet. Traditionally, keyword extraction always depends on the word dictionary and word frequency. However, it can not extract the keywords out of the dictionary and low-frequency keyword such as Chinese name, company's name, and so on. In this paper, we propose several new techniques to do keyword extraction efficiently for Chinese Web page. We get Web pages dynamically and do segmentation by Viterbi algorithm. Then we use some strategies to find those keywords that are not included in the Mandarin dictionary. Experiments show that both the precision and recall rates are satisfied. And we will compare our system with the other system.

1. INTRODUCTION

One important feature of Chinese texts is that they are character-based, not word-based. There are no blank to mark word boundaries in Chinese text. In the Internet, there are too few corpus in one Web page. So it is difficult to extract keyword from Web pages by traditional statistically method. There are some researches concerning about keyword extraction technique. For instance, Keh-Jiann Chen et al.[1] proposed a method based on the prefix-category and suffix-category associations in the prediction of the pos-categories of Chinese unknown word. Lee-Feng Chien et al.[2][3],their method are based on modified Mutual-information and uses PAT-Tree data structures to reduce the computation time. Another method[4] that iteratively integrates the contextual constraints and a joint character association metric to progressively improve the segmentation results of the input corpus was proposed by Jing-Shin Chang et al. Hsin-min Wang et al.[5], their approach segments Chinese Web page's hyperlinks and bookmarks based on a Chinese lexicon, those remaining single-character strings are treated as a dynamic keyword lexicon.

For keyword extraction, the unknown words are defined as the words which are not in the lexicon. The following types of unknown words often occur in the corpus.

(1). Chinese names:e.g. "黃介一(Huang Chieh-Yi)"
(2). Company and Agency name: e.g., "國立成功大學"
(3). Abbreviation words: e.g., "台汽(Taiwan-bus)", "中油(China-fuel)".
(4). New words:"辣妹(Spicy-girl)".
(5). Numerical strings:"1998 年(1998-yuear)".

In this paper, we propose several methods to extract these kinds of unknown words. First, we use the traditional statistically method to do segmentation and extract keywords that in the Mandarin dictionary. Then we will describe the other methods and our system in Section 2. Section 3 is the experimental results are presented. And finally, Section 4 gives the conclusion.

2. SYSTEM DESCRIPTION

The architecture of our system is shown in Figure 1. We use relaxation strategies to do keyword extraction. In other word, we try to find all the possible keywords first (Phase 1). Then we refine these keywords by some rules (Phase 2).

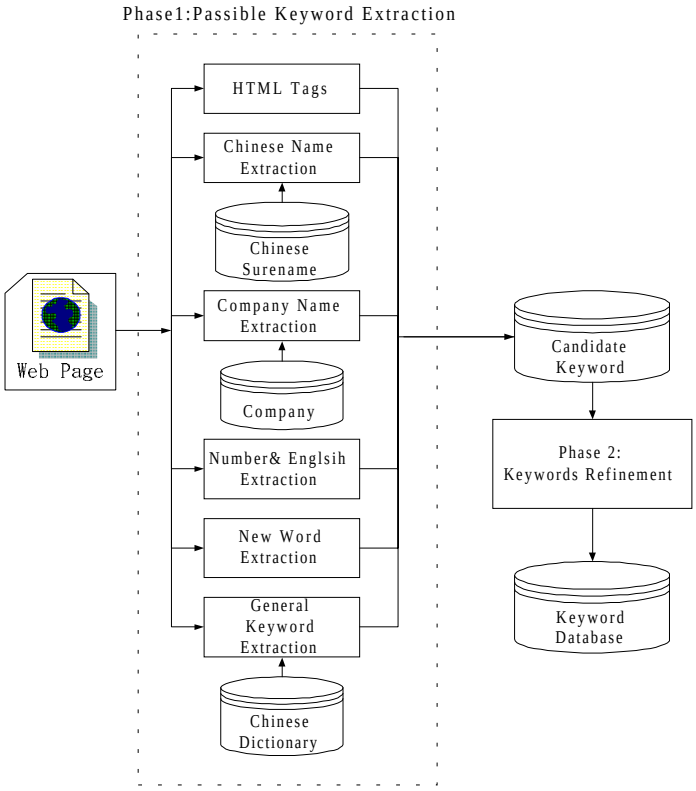

Figure 1. The system architecture of Keyword extraction system

Phase 1. Possible keyword extraction..

First, the words which can be found in the dictionary are defined as keywords. Then we develop some methods to find the following keywords.

- HTML tags: We can extract the keywords by some HTML language tags and Hyper-Link, such as, "<title>", "<H1>", "<H2>", ...,"<H6>","<B>", and so on. The words tagged by these tags are defined as keywords. For instance, <title> 實驗室 (Laboratory)</title>, <B>工學院(College of Engineering) </B>.
- Chinese name: Most of Chinese names consist of one-character surname and one to two-characters first name. In Chinese, there are about 100 surnames(Table 1) totally. We can find surname using table search then add one to two characters after surname to form the two possible Chinese names. e.g., “王章(Wang Chang)”, “黃介一(Huang Chieh-Yi)”.

Chinese surname
趙,錢,孫,李,周,吳,鄭,王,馮,陳,魏,蔣,沈,韓,楊,朱,秦,尤,許,何,呂,施,張,孔,曹,嚴,華,金,衛,陶,姜,戚,謝,鄒,柏,章,雲,蘇,潘,葛,范,彭,魯,韋,馬,苗,鳳,花,方,俞,任,袁,柳,鮑,史,唐,費,岑,薛,雷,賀,倪,湯,殷,羅,畢,郝,傅,皮,卞,齊,康,伍,余,卜,顧,孟,黃,穆,蕭,尹,姚,邵,汪,祈,毛,狄,米,貝,成,戴,談,宋,龐,熊,紀,屈,項,祝,董,梁,杜,阮,藍,席,季,麻,賈,江,童,顏,郭,梅,盛,林,刁,鍾,徐邱,駱,高,夏,蔡,田,樊,胡,凌,霍,虞,萬,柯,管,盧,莫,經,房,裘,繆,干,解,應,宗,丁,宣,鄧,郁,單,洪,包,左,石,崔,鈕,龔,程,邢,裴,陸,翁,羊,封,靳,井,段,富,巫,烏,焦,巴,牧,谷,車,侯,全,郗,班,秋,仲,伊,宮,仇,甘,厲,戎,祖,武,劉,詹,龍,葉,黎,白,賴,卓,蘭,屠,蒙,喬,申,璩,桑,桂,牛,農,溫,莊,晏,柴,瞿,閻,艾,魚,容,向,古,易,戈,廖,庾,耿,文,寇,歐,師,鞏,庫,聶,勾,敖,融,冷,那,簡,曾,養,相,查,游,權,張簡,歐陽,司馬,上官,夏侯,諸葛,東方,尉遲,公羊,公冶,單于,申屠,公孫,軒轅,令狐,慕容,司徒

Table 1. 100 Chinese surnames.

- English and number: Chinese web page may also contains English words. We collect all the English words in the web to be keywords. The numbers are defined as keywords for the same result. e.g., “1998 年(1998-year)”,
- Company and agency name: Firstly, we find the words corresponding to the company or agency, such as "company", "Co.", "university", "high school", "hotel", "government", and so on(Table 2). Then we add two to four characters before these words to form the keywords. e.g., “中國石油公司”, “國立成功大學”.

Company and Agency
公司,集團,行,號,小學,國中,中學,高中,大學,有限公司,飯店,旅社,博物館,美術館,館,組,局,站,分局,室,處,辦公室,科,所,研究所,中心,推廣中心,研究室,課,中心,小組,指揮中心,派出所,隊, 工業區,工廠,場,廠 ... etc.

Table 2. Company and agency

- New word or Abbreviation words: We use statistical method to find the new words and abbreviation words.

	Method	Need dictionary	Able to extract			Precision Rate	Recall Rate	Using on WWW
			Proper name	Abbreviation	Compound			
Chang Jing-Shin [4]	Unsupervised Iterative and Likelihood Ratio Ranking Model	√	√	High frequency words		68%	76%	Not adaptation
Chen Keh-Jiann [7]	1.segmentstion 2.tagging 3. Using rule	√	If match rule			74% rule>75%	74% rule>75%	Not adaptation
Chen Hsin-Hsi [8]	Using spelling method, adjacency principle and HTML tags.	X	√	In match rule	N/A	43%	84%	Just using in WWW
Shyuu W.L. [9]	Combine Statistic, PAT-tree and Rule	√	√	√	√	N/A	96% unknown 45%	Not adaptation
Lin Yih-Jeng [10]	Using occure times and refinement.	X	High frequency words			N/A	N/A	Not adaptation
Wang Hsin-min [5]	Segmentation and combine remaining single-character strings	√	√	In match rule	X	N/A	N/A	Just using in WWW
Our system	Proposed in this paper	√	√	Hgh frequency words	√	75%	>90%	√

Table 3. Compare table.

Phase 2. Keyword refinement

Since we use the relaxation strategies, there will be too many keywords obtained in phase 1. In order to discard the meaningless keywords, the following strategies are used to refine the keywords.

- The keywords in the dictionary should not be separated. e.g., “開張大吉”, Since “張” is a Chinese surname, “張大吉” or “張大” maybe a Chinese name. But “開張大吉” in our dictionary so we discard “張大吉” or “張大” in our candidate keyword.
- Using the grammar rules to refine the keywords.
  1) Using the template to refine the keywords. For instance, “國立 XX 大學(National XX University)”. The name of university, XX is chosen as a keyword, but candidate keywords 立 XX and 國立 XX are discarded.
  2) When extracting Chinese name from a sentence, the segmentation result which contains less signal keywords is prefered. For instance, “李來發明天去台北” there are two results, “李來 發明 天 去 台北”, “李來發明天 去 台北”. So we choose the second sentence and discard “李來”,“發明”.
  3) If the words before a company name are adjs or VP (such as “來”, “去”, “到”, “的”, “這個”, “那個”, etc.) then these words are not keywords and are discarded.
- Using bigram information to discard the keywords of low probability.

3. EXPERIMENTAL RESULT

To check the result of our system, we randomly choose some Chinese homepages for processing. We determine keywords manually. Then our system extracts keywords for the same webs automatically. These two sets of keywords are analyzed to know which keywords are lost or how many meaningless keywords are mis-extracted. In addition, some comparisons are made between our approach and the existing methods. We test several Homepages and the performance of experimental results are satisfactory. In Table 3, we compare our system with the other keyword extraction system.

4. CONCLUSION

In this paper, we develop several new techniques to take into account the characteristics of Chinese language for the word extraction in Chinese web pages. The experiment shows that the proposed approach is feasible and efficient. In the future, we will extend our system to be a web understanding system.

REFERENCES

[1] Keh-Jiann Chen et al., “Category Guessing for Chinese Unknown Words”, SNLP’97 pp. 35-40

[2] Lee-Feng Chien et al. "Networked Chinese Information Access Using Speech and Natural Language Information Retrieval Techniques", ICCPOL'97, pp. 669-674

[3] Lee-Feng Chien, “PAT-Tree-based Keyword Extraction for Chinese Information Retrieval”, 1997 ACM SIGIR Conf. On R&D in IR, pp.50-59

[4] Jing-Shin Chang and Keh-Yih Su, "An Unsupervised Iterative Method for Chinese New Lexicon Extraction", Computational Linguistics and Chinese Language Processing vol. 2, no. 2, pp. 97-148. August 1997

[5] Hsin-min Wang et al. "Surfing the Chinese Web Pages by Unconstrained Mandarin Speech”, ICCE’98, pp. 84-85.

[6] Yuen-Hsien Tseng, "Fast Keyword Extraction of Chinese Documents in a Web Environment”, IRAL'97.

[7] Keh-Jiann Chen and Ming-Hong Bai, "Unknown Word Detection for Chinese by a Corpus-based Learning Method", Computational Linguistics and Chinese Language Processing vol. 3, no. 1, pp. 27-44. February 1998.

[8] Hsin-Hsi Chen and Guo-Wei Bian, "While Page Construction from Web Pages for Finding People on the Internet", Computational Linguistics and Chinese Language Processing vol. 3, no. 1, pp. 75-100. February 1998.

[9] Chen C.C., Shyuu W.L.,"A Multi-layer Chinese Segmentation System with Combined Statistic and Rules”, ROCLING XI '98. pp. 63-72

[10] Yih-Jeng Lin et al. “A Way to Extract Unknown Words Without Dictionary from Chinese Corpus and its Applications”, ROCLING XI '98. pp. 217-226