Workshop on Artificial Intelligence

A New Two-Stage Keyword Extraction Method for T.V. News Contents

T. M. Fang, C. L. Tseng, H. C. Fu
Department of Computer Science and Information Engineering
National Chiao Tung University, Hsinchu, Taiwan, 300, ROC
E-mail: { tmfang, tcl, hcfu }@csie.nctu.edu.tw
Tel: 03-5731930
Corresponding Author: Professor H.C. Fu

Abstract

In this paper, we address a two-stage keyword extraction method for Internet Chinese T.V. news document. The first stage, a lexicon matching process, extracts most meaningful words from a document. The second stage performs a statistical analysis of the term frequency of each word in the document. We also propose an integrated algorithm to combine these two results to conclude a keywords list that can best outline or represent the original documents. We have implemented a prototype system to evaluate the functionality of the proposed method. Preliminary experimental results show that the proposed method seems very promising. We also implement this keyword extraction and indexing methods in a Internet TV news browsing system, http://nn.csie.nctu.edu.tw/Project1-1/introduction.htm. A user can give set of query keywords to retrieve preferred multimedia TV news stories.

Keywords: Chinese word segmentation, keyword extraction, lexicon matching, statistics analysis, Internet T.V. news

A New Two-Stage Keyword Extraction Method for T.V. News Contexts

Abstract

In this paper, we address a two-stage keyword extraction method for Internet Chinese T.V. news document. The first stage, a lexicon matching process, extracts most meaningful words from a document. The second stage performs a statistical analysis of the term frequency of each word in the document. We also propose an integrated algorithm to combine these two results to conclude a keywords list that can best outline or represent the original documents. We have implemented a prototype system to evaluate the functionality of the proposed method. Preliminary experimental results show that the proposed method seems very promising. We also implement this keyword extraction and indexing methods in a Internet TV news browsing system, http://nn.csie.nctu.edu.tw/Project1-1/introduction.htm. A user can give set of query keywords to retrieve preferred multimedia TV news stories.

Keywords: Chinese word segmentation, keyword extraction, lexicon matching, statistics analysis, Internet T.V. news

I. Introduction

Due to the advances of the network technologies, more and more multimedia documents are available on the Internet. Automatic and efficient analyzing these documents so as to mining some valuable information often needs a powerful indexing system. Keyword indexing is definitely an efficient tool among the various tools and methods for indexing. In this paper, we address a two-stage keyword extraction method for Chinese document available on Internet T.V. news. However, the layout of Chinese characters in a sentence is different from the layout of alphabetical words in spelling languages. Particularly except the character spaces, there is no extra space or symbols between two Chinese characters within a sentence.

Hence, it is a difficult task to correctly extract meaningful words from a sentence. In addition, it is more difficult to pick out the most representative keywords from a document.

Although it seems difficult, many methods have been proposed to efficiently perform word segmentation, and further more to extract keywords from Chinese documents. For example, [1] uses lexicon together with Chinese word construction rule to eliminate improper segmentation results in order to find a correct one. Sporat et al. [2] employed a statistical method to group Chinese characters into twocharacter words making use of a measure of character association based on mutual information.

Chien [3] proposes a method that maps a document into PAT-Tree's form, and then calculates the term frequency for each string (or word) to eliminate the words which are statistically incomplete. Therefore, the remaining words are decided as candidates of Significant Lexical Pattern (SLP). They also applied the general lexicon to delete some most frequently appeared words, which are often non-meaningful, thus the remained words that can truly and efficiently represents or outlines a document. Because the PAT-Tree is similar to binary search tree in structure, therefore it is convenient and efficient to construct a new node and also to search a string in O(logN)

time.

In addition to the keyword extraction methods on general document, there are several different keyword extraction methods used by some Internet search engines. Search engines extract keywords for the purpose of efficiently indexing an Internet document. For example, Google[4] uses the method that adjusts the keyword weight according to the location of the keyword in the document to emphasize the importance among specific HTML tags. These methods primarily focus on Internet documents in HTML syntax form, and it may not be suitable for processing general non-HTML documents. In the paper, we are interested in the processing of non-Marked documents. We propose a two-stage method for Chinese words segmentation. The two-stage are (1)lexicon matching and (2)statistics analysis. By using this method, the most representative words are selected by using the lexicon matching, and the most frequently used words are selected by statistical analysis. Therefore, the extracted keywords of a document contains (1)representative words, such as person, location, and organization names and (2)most frequently used words which may not be included in (1). Our preliminary results (see Section III.D) show that this method produces promising results.

In the rest of this paper, the detail of the proposed method will be described in Section II. Experimental system and preliminary results will be provided in Section III. Also, the quality of extracted keywords will be discussed. Finally, we will draw concluding remarks and future work in Section IV.

II. Lexicon Matching based Keyword Extraction

In general, there are two major types of method for Chinese word segmentation. One is statistics analysis method, and the other is lexicon matching method. Statistics analysis method calculates the statistics of the term frequency of strings in a document, and uses a pre-defined analysis model to find the best fit segmentation results, and then term frequency are used to select suitable terms as keywords. Lexicon matching method searches all the words in the lexicon to match the strings in the document in order to find all the possible words. Thus, the keywords are selected according to the property of the word in lexicon.

Statistics analysis method has been widely used because it can overcome some limitations of the lexicon method. For instance, an often happened problem that new or specific words are usually not included in the lexicon, and the processing speed of

statistical method is faster than lexicon matching method. However, there still exists some disadvantages of statistics analysis method. The keywords extracted by statistics analysis method may only have the key property in statistical sense, and no one can be sure that these keywords also have some key properties or meanings in semantics. Therefore, the capability of using these keywords to outline or to index a document can be very limited.

By including the additional lexicon matching method, extract grammatical properties of each word can be specified, so that selected words can best describe a document. For instance, the five ‘W’ elements, e.g., Who, What, When, Where, How are key properties in a news documents. To be more specific, we propose to use the lexicon matching as a major tool, and the statistics analysis as a helping tool. In other words, extracting the keywords are first matched or searched in a lexicon, then select the other keywords which are outside the lexicon but are with significant statistical property. And these keywords are also the candidates waiting to be added into lexicon. Hence, the semantic properties which are provided by lexicon and the statistical property such as term frequency which are provided by statistics analysis would produce sufficient information in selecting a keyword from a document, and it also make the keywords extracted with higher confidence level in quality.

A. Lexicon Selection

In Chinese text, a word may play different roles depends on its context in the sentence. Thus a word may represent different meanings in consequence. Therefore, it is necessary to check the completeness and generality when using the lexicon, that means the lexicon should contain the most usage of a word with different context (completeness), and it should also contain the most general words in different classification domain of the daily life (generality).

Based on the analysis on the above two constrains, we selected and used the lexicon developed by CKIP (Chinese Knowledge Information Processing)[5] group in Academia Sinica. This lexicon contains over 80 thousands of words, and each word is documented with (1) statistical property such as term frequency, which are calculated from source corpus, (2) syntax classification and the semantic tags of some syntax classes. The syntax classification is performed by human according to the ICG (Information-based Case Grammar) developed by CKIP. By adding with semantic tags, document of the syntax classification contains sufficient information, thus it is suitable for language analysis, which is a key requirement for us to select this lexicon

in our system.

B. Integrated Algorithm for the two-stage keyword extraction methid

The flow diagram of the integrated algorithm is shown in Fig. 2-1. Overall, the integrated algorithm contains four parts.

Part 1. Construct PAT-Tree and Select SLP (Significant Lexical Patterns)

At the beginning, the document is mapped on to a PAT-Tree form, such that important words can be represented with different statistical properties. There may be some new or specific words outside a lexicon, this step calculates the statistics of these words, to prevent them from being included in the keywords. A measure called SE (Significant Estimation) was used here to indicate what are the most significant words in the PAT-Tree. Since the concept of SE was based on the substrings' mutual information, thus, the words in a PAT-Tree can be sorted according to their SE value, such that the best keyword group to describe the document can be selected and the words that are no so representative will be rejected by a pre-defined threshold.

Part 2. Check Non-Single Character Word

According to a general survey and observation, we found that the length of most of Chinese words is normally less than or equal to four characters, if we use the general lexicon mentioned in Section II.A as sample space, then there are 98.48% words with length less than or equal to four characters. Besides, we found that

Lexicon Matching based Keyword Extraction Flowchart

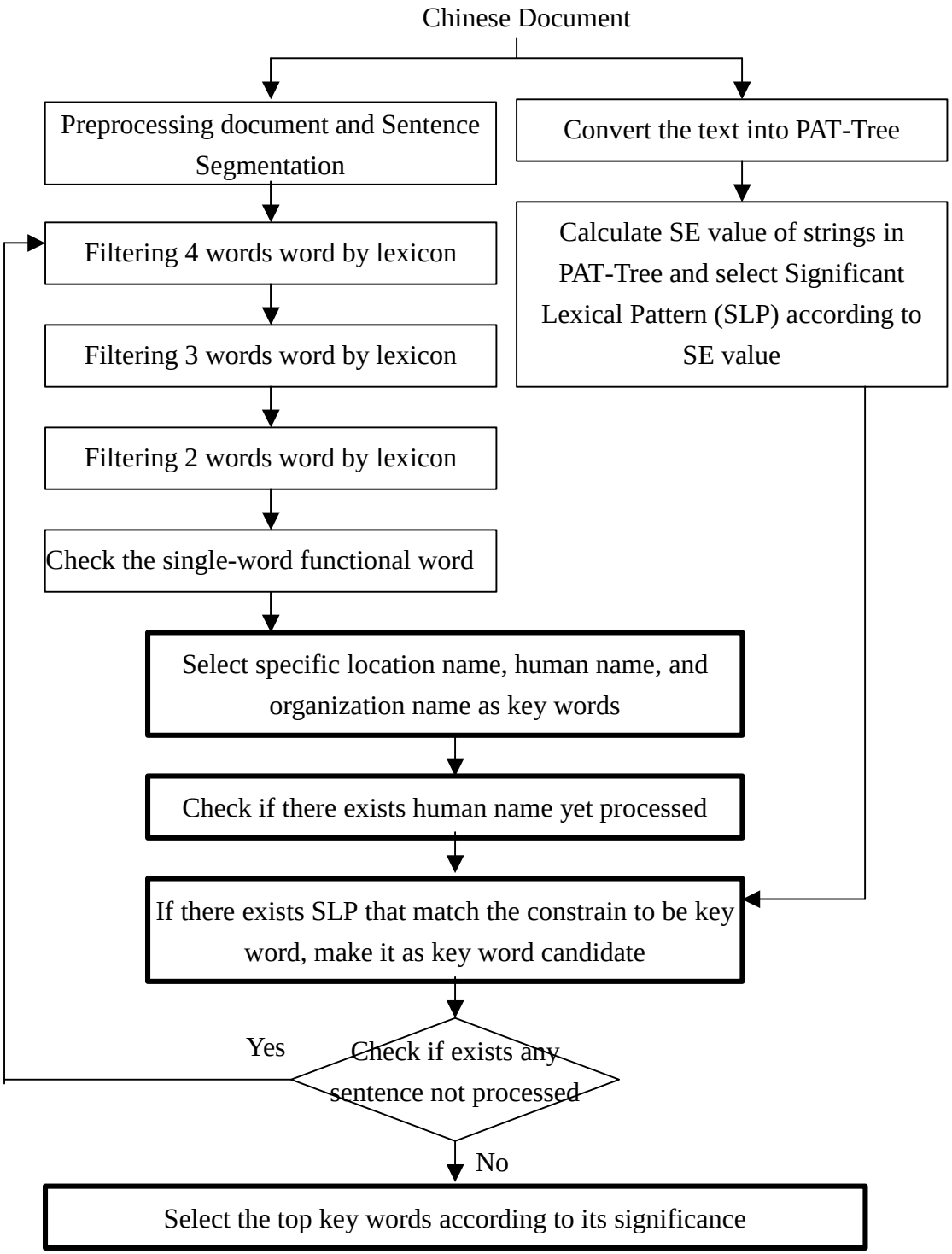

Fig. 2-1: The flow diagram of the integrated algorithm for the two-stage keyword extraction method.

Chinese characters have the tendency to form a meaningful word with longer length. In other words, if a two characters word is contained in a four characters word, then the four characters word is usually the most meaningful segmentation in the sentence.

Consequently, our method would not consider the words of length more than four characters to save some processing time, and only search the unknown string in the lexicon of the order from four characters word to two characters word in order to follow the tendency that longer length word dominates in sentences.

Part 3. Select Functional Single Character Word

After the process of checking non-single character words on a sentence, groups of meaningful words are selected and some single character are left over. On these remained characters, we would like to find out the functional single character which does not have any semantic meanings. These characters are often to be preposition, conjunction, adverb, and pronoun, etc. These characters could not be one of the five 'W' elements, but it could occur with other meaningful words such as noun and verb to make description more detail with subject. In general, we would not select these words as keywords, but these functional characters could still provide reference information for deciding a word extracted from PAT-Tree to be keyword or not. Hence, these functional single-character words are still selected and will be used later.

Part 4. Select the Keywords

In this part, we would combine the results generated by the previous processes to make a total solution in selecting the most representative keywords of the document. In general keyword extraction method of statistics analysis, it usually uses the concept of TFIDF (Term Frequency Inverse Document Frequency)[6] as a measurement to decide the significance of words. In fact, TFIDF means if a word occurred in a document many times and only existed in a few document, then we thought this word is more representative for this document than other words. However, as we mentioned in the beginning of this section, the keywords for a news document always needs the five 'W' elements, Who, What, When, Where, and How to be the very basic keywords. Therefore, the TFIDF can not be best fit to our requirements, we proposed the following series of filters to select the keywords to fulfill our need.

Type I filter: Proper Nouns in the lexicon

Specific location, person, and organization names have their unique class in the lexicon respectively. That is to say, we can select these words according to the class in lexicon to help us find the elements of Where and Who.

Type II filter: Person's Name

Usually, person name would not be appeared in the lexicon if he or she is not famous enough. However, due to the list of Chinese Family Names (百家姓), we can check the possible human names in sentences more easier. After the process of lexicon searching, if there exists a series single characters of total length equal to or less than four characters, that means no known word is associated with them, and also the heading character is in the list of Chinese Family Names, then we could conclude that these single characters might be a person's names. To check if these characters are truly names, we need to check the character or word right before and after the possible name string, to see if the combination of them matches the rule of Chinese grammar. When all these checking are passed, we could claim this string as a candidate person's name.

Type III filter: SLP (Significant Lexical Pattern) outside the Lexicon

Due to the size limitation of a lexicon, some proper nouns such as person's names or organization names would not be included in a lexicon. Thus, when a SLP is extracted according to its SE (Significant Estimation) value from the PAT-Tree, this SLP might be a keyword outside the lexicon, and in most cases, it is also a proper noun. In this situation, we need to group the SLP and its context as a whole to see if they are in the correct order and follow the correct rules of

sentence construction according to Chinese grammar. If this SLP passes all these checks, then this SLP can be a complete full meaning keyword.

Type IV filter: SLP within the Lexicon

Suppose a SLP is contained in a lexicon but it is not one of the above three types, this SLP could be a functional word, or a possible keyword. By using the method proposed in Part 3, we can expel the SLP to be a functional word. Then, each remained SLP is calculated for its Representative Estimation (RE) value according to the statistics in the lexicon. Basically, the RE value is derived from the similar concept as TFIDF. Hence, we could use the RE value to sort the SLPs to find the most suitable keywords to describe the document.

By using the proposed four steps, some keywords can be extracted from a document, and these keywords could be used to retrieve further more information or some data mining works.

III. Experimental Results

A. System Overview

The proposed keyword extraction system structure is shown as Fig. 3-1. This system

collects news text available on the Internet, and then extracts sets of keywords from each news documents according to the proposed method. Each set of keywords will be inserted into a TV news database for later processing, such as data mining or information retrieval. A user can browse and query desired news text by giving a set of keywords through an interactive interface. We will use this interface to perform some experiments to evaluate the proposed keyword extraction system.

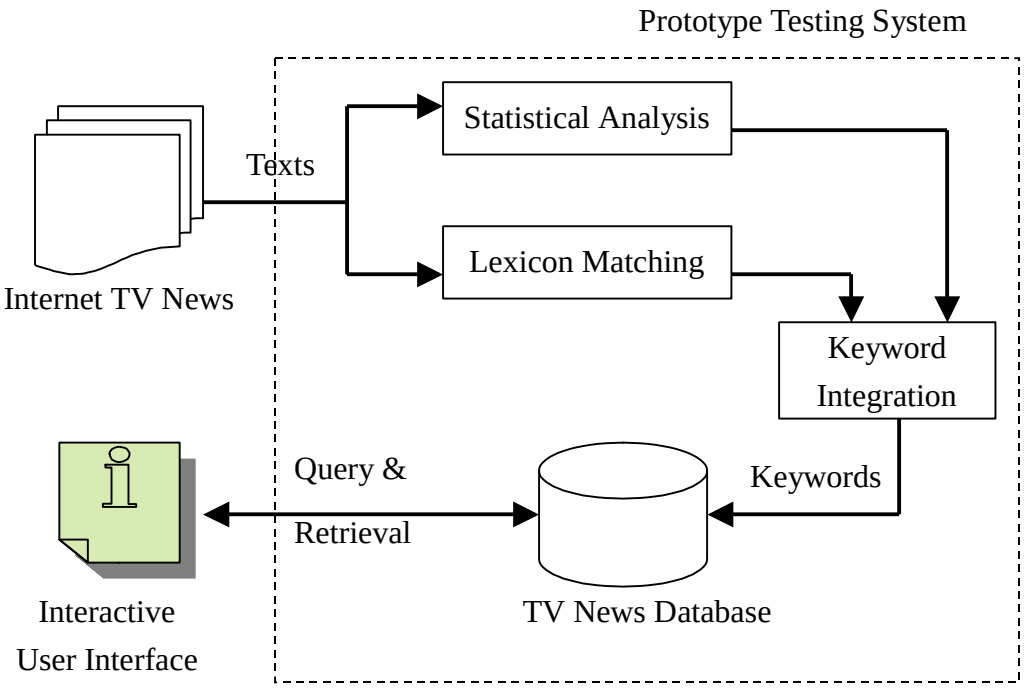

Figure 3-1: The proposed two-stage keyword extraction testing system.

B. Testing Database

Since there is no standard database and/or benchmark to evaluate the quality of the keyword extraction process, the following keyword evaluation may be very subjective. In our experiment, we proposed a method that the extracted keywords are

evaluated according to the evaluation criterion by four persons, then the averaged score of each text is considered as the quality of the keywords. Due to the time limitation to prepare this paper, we only randomly select five days in July, 2002, and collect the news text from the web site of CTS (中華電視) of these days to be testing database.

C. Evaluation Parameter

As shown in Table 3-1, a scoring rule is defined for extracted keyword evaluation from a text. The lowest score means that extracted keywords is not related, and the highest scores indicates the proper nouns (such as person, location, and organization names) and key-concepts are all collected in the keywords.

Table 3-1：Keyword Evaluation Score

Score	Score Description
3	Proper nouns and significant concepts are included
2	Proper nouns are included, but some concepts are redundant or lost
1	Some proper nouns are lost, and concepts are redundant or lost more
0	Keywords extracted are not related to the news text

D. Preliminary Results and Discussion

After the scoring, we get the preliminary result as Table 3-2.

Table 3-2：Experimental Results

Date	Average Score	Standard Deviation
7/03	2.000	0.655
7/04	1.933	0.727
7/05	2.103	0.759
7/11	2.100	0.597
7/12	1.688	0.726

Preliminary results show that keywords extraction is very promising in average, and we can find the representative keywords in most texts. However, standard deviations are large in contrast to average scores, that means the quality of keywords extraction seems floating up and down around an acceptable average score line.

In the process of selecting keywords, some parameters could be adjusted to make the quality of automated keyword extraction to be near to professional level. For example, when a SLP is selected from a PAT-Tree, a threshold TH_{SE} is used to filter out candidate words with low confidence. Therefore, TH_{SE} could be adjusted according to the experiment results to improve the quality of keywords extraction.

In the meantime, we have embedded the keyword extraction and indexing subsystems in an Internet TV news system, http://nn.csie.nctu.edu.tw/Project1-1/introduction.htm. The TV news system constantly converts TV news program into an Internet multimedia TV contents. A user can select a date to browsing daily TV news program in text, key-frame images, and story based news video. As shown in Figure 3-2, by

entering a keyword, six related news stories are retrieved from the Internet TV News system.

Figure 3-2: By entering a query keyword, a user can retrieve six related news stories from the Internet TV News system.

IV. Concluding Remarks

In this paper, we have proposed a new two-stage keyword extraction method. The proposed method combines both advantages from lexicon and statistic methods and complement each other's disadvantages. Preliminary experimental results show that the two-stage method seems very promising. We also implemented the keyword

extraction and keyword indexing subsystems in an Internet TV news system. Thus, a user can retrieve his/her favored or preferred TV news stories by giving a few query keywords.

References

[1] 陳克健, 陳正佳, 林隆基, "中文語句分析的研究-斷詞與構詞", *中央研究院資訊所技術報告 TR86-00*4, 1986.

[2] R. Sporat, C. Shih, "A Statistical Method for Finding Word Boundaries in Chinese Text", *Computer Processing of Chinese and Oriental Languages*, Vol. 4, No. 4, pp.336-351, 1990.

[3] L. F. Chien, "PAT-Tree-Based Keyword Extraction for Chinese Information Retrieval", *the 20th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval Philadelphia*, pp.50-58, PA, USA, July, 1997.

[4] Google, "http://www.google.com", Search Engine.

[5] CKIP, http://godel.iis.sinica.edu.tw/CKIP/, 中文詞知識庫小組.

[6] Gerard Salton, "Automatic Text Processing", *Addison-Wesley*, December, 1988.