Decryption of Full Text Retrieval Technology: Chinese Word Segmentation

Xuebing Lu[1, a], Yili Xu[1, b], Weiwei Deng[1, c], Yingjie Yan[1, d]

[1] Shanghai Entry-Exit Inspection and Quarantine Bureau, Pudong New Area, Shanghai, China, 200135

[a] email, [b] email, [c] email, [d] email

Keywords: Segmentation Method, Recognition, Chinese Word Segmentation

Abstract. Based on the development of full text retrieval function of administrative office system of Shanghai Entry-Exit Inspection and Quarantine Bureau, this paper comprehensive introduces the Chinese segmentation technology used in full-text retrieval. The three mentioned methods, which are segmentation method based on string matching, the segmentation method based on comprehension and the segmentation method based on statistics. The advantages and disadvantages of the three segmentation methods are compared in this paper. The two difficult points of ambiguity recognition and new word recognition are also discussed in the paper.

Introduction

The system stability, security and convenience of the administrative office system of Shanghai Entry-Exit Inspection and Quarantine Bureau (hereinafter referred to as the "Shanghai bureau") has been greatly improved since the upgrade to the B/S architecture of 2013. The current administrative office system as the core system of Shanghai bureau chief platform, set a business function dispatch module, file management module and the financial management module, performance module, quality system module, electronic statements, 2015 will be the original OA system in the history of the dispatch system is imported, gathered a massive data, according to the background database to calculate, the unstructured data occupies 80%, reached about 60G, currently provides administrative office system query is the data retrieval based on, it is difficult to meet the requirements of precision and speed of retrieval, especially the retrieval based on the content of the document. In order to provide users with pain points, the information is in the development of this year for the full text retrieval of the administrative office system.

Full text retrieval technology is a branch of information retrieval. Chinese based on search engine technology provides a new and powerful search function of Shanghai Bureau of full-text retrieval, not only can realize all functions of the traditional literature retrieval, but also directly according to the data retrieval system. This paper will be one of the core contents of the full text retrieval technology: Chinese word segmentation.

What Does the Chinese Word Segmentation Mean?

As is known to all, English is a unit of words, words and words are separated by spaces, and Chinese is in the word as a unit, all the words in the sentence can be linked together to describe a meaning. For example, the English sentence am a teacher I, in Chinese, as: "I am a teacher". The computer can be very simple to know that teacher is a word, but cannot easily understand the

"teaching", "division" two words together to say a word. The sequence of Chinese characters segmentation Chinese into meaningful words is Chinese word, some people also called segmentation. I am a teacher, the result of word segmentation is: I am a teacher.

Technology of Chinese Word Segmentation

Segmentation Method Based on String Matching. This method is also called mechanical word segmentation method, which is in accordance with certain strategies to be analyzed and Chinese characters on a "sufficiently large" machine dictionary entries are supported, if find a string in the dictionary, there is a match. According to the scanning direction, string matching segmentation method can be divided into positive matching and reverse matching; according to different length matching, can be divided into the largest (longest), and minimum (shortest) matching; according to whether the POS tagging process and combination, and can be divided into the integration method of simple word segmentation and word segmentation with the combination of annotation. Several common mechanical word segmentation methods are as follows:

(1) Maximum Matching Method. This method is also known as the 5 - 4 - 3 - 2 - 1 query method, the basic idea is: the assumption that the longest term automatic dictionary contains a number of M Chinese characters, then take the material to be processed in the current string number in M characters as the matching field, find the dictionary, such as the existence of a m such words in the dictionary, there is a match, the matching field as one word is cut out; as to a M word that was not found in the dictionary, it will remove the last matching field Chinese characters, the remaining M-L characters as matching new fields, new matching, and so forth a word segmentation, until the completion of a match, cutting out a word, according to the above steps, until all the word segmentation.

(2) Reverse Maximum Matching Method. With the above method is different from the beginning of the sentence at the end of the treatment, each time the match is not successful removed the most front of the field of a Chinese character. The experiments show that the RMM method is more accurate than the MM method.

3) Minimum Segmentation Method. The method that based on Omni segmentation, find a minimum segmentation words in a sentence sum. Various methods can be combined with each other, for example, a bidirectional matching method can be formed by combining the forward maximum matching method and the reverse maximum matching method. Because the Chinese word characteristics, positive minimum matching and reverse minimum matching generally rarely used. Generally speaking, the segmentation accuracy of the reverse matching is slightly higher than that of the forward matching, and the ambiguity is less. Statistical results show that the simple use of the maximum matching error rate is 1/169, the error rate of the simple use of the reverse maximum matching is 1/245., but the accuracy is still far from meeting the actual needs. The actual use of the word segmentation system, are the mechanical word segmentation as a primary means, but also through the use of a variety of other language information to further improve the accuracy of segmentation. One way is to improve the scanning mode, called feature scan or symbol segmentation, priority identification and cutting out some words with distinct characteristics to be analyzed in a string, with these words as the breakpoint, the original string is divided into smaller strings come into mechanical segmentation, thereby reducing the matching error rate. Another method is to combine segmentation and lexical category labeling, help to make full use of the segmentation decision lexical category information, and in the process of marking in turn on segmentation results for inspection, adjustment, thus greatly improve the accuracy of segmentation.

Segmentation Method Based on Comprehension. This kind of word segmentation method is through the computer simulation of human understanding of the sentence, to achieve the effect of the word recognition. The basic idea is to use syntactic and semantic analysis in Chinese word segmentation, and to deal with ambiguity by using syntactic information and semantic information. It usually consists of three parts: word segmentation system, French justice system, the total control part. In the coordination of the total control part, the word segmentation subsystem can obtain the syntactic and semantic information of the word, sentence and so on to judge the ambiguity of word segmentation, that is, it simulates the process of human understanding of the sentence. This method requires the use of a large number of language knowledge and information. Because of the generality and complexity of Chinese language knowledge, it is difficult to organize all kinds of language information into the form of the machine can be read directly, so the present system of the word segmentation based on understanding is still in the experimental stage.

Segmentation Method Based on Statistics. From a formal perspective, the word is a combination of a stable word, so in the context of the number of adjacent words at the same time, the more likely to form a word. Therefore, the frequency or probability of adjacent co-occurrence of words and words can reflect the credibility of the word. We can count the frequency of the combination of each word in the data, and calculate their mutual information. Define the mutual information of two words, and calculate the probability of two Chinese characters Y and X. Mutual information embodies the close relationship between Chinese characters. When the close degree is higher than a certain threshold, it is thought that the word group may form a word. This method only needs a statistical word frequency in the corpus, without segmentation dictionary, which is also called no dictionary lexical or statistical check method. But this method also has some limitations, often taking some co-occurrence frequency is high, but not the word commonly used word, such as "this", "one" and "some" and "I" and "many", and the commonly used word recognition accuracy, time and space overhead big. The statistic system in practical application is to use a basic word dictionarystring matching, and use statistical methods to identify some new words about string frequency statistics and string matching together, both play and segmentation to match the characteristics of fast speed and high efficiency, and the no context identification of new words, to automatically disambiguate the advantages of dictionary segmentation.

Difficulties in Chinese Word Segmentation

Ambiguity Recognition. Ambiguity refers to the same sentence, there may be two or more of the segmentation method. For example: the surface, because the "surface" and "face" are words, so the phrase can be divided into "surface" and "surface". This is called cross ambiguity. Such cross ambiguity is common, such as "make up and dress" can be divided into "make up and clothing" or "make up and dress". Cross ambiguity is relatively easy to handle, and the combination of ambiguity must be judged according to the whole sentence. For example, in the sentence "the door handle is broken", "handles" is a word in the sentence, but please "hands off", "handles" is not a word; in the sentence "the appointment of a lieutenant general", "will" is a word, but the sentence "the output three years, an increase of two times", "will" is no longer a word. How do these words computer to identify? If the computer can solve the problem of cross ambiguity and combination ambiguity, there is still a difficult problem in the ambiguity. For example: "badminton" can be cut into the auction is over, "badminton", also finished auction may cut into the "auction". If there is no badminton, other context sentences, I am afraid that no one knows "auction" here is not a word.

Table 1. Advantages and Disadvantages of the Three Segmentation Methods

Segmentation Method	Segmentation Method Based on String Matching	Segmentation Method Based on Comprehension	Segmentation Method Based on Statistics
Ambiguity Recognition	bad	good	good
New Word Recognition	bad	good	good
Need Dictionary	yes	no	no
Need Corpus	no	no	yes
Need Rule Library	no	yes	no
Algorithm Complexity	easy	difficult	normal
Technology Maturity	maturity	immaturity	maturity
Difficulty of Implementation	easy	difficult	normal
Accuracy of Segmentation	normal	accurate	relatively accurate
Speed of Segmentation	fast	slow	normal

New Word Recognition. New word, professional term is called the unknown word. That is, those in the dictionary are not included, but really can be called the words of those words. The most typical is the name of a person, people can easily understand the sentence "Wang Junhu to Guangzhou", "King" is in a word, because it is the name of a person, but if the computer is difficult to identify. If the king "in" as a words to the dictionary, there are so many names around the world, and there are always new names. In addition to new names, and names, place names, product name, brand name, abbreviation. At present, the new word recognition accuracy has become one of the important marks of the evaluation of a word segmentation system.

Conclusion

Chinese segmentation emerged as the times require. It can meet the needs of information search and solve some communication obstacles between person and computer to a large extent. Through the in-depth research on Chinese segmentation technology, we believe that we will develop high quality and multi-function segmentation systems to promote the wide application of Chinese information filtering system in the next few years.

References

[1] Wang Liwang, Journal of Southwest China Normal University (Natural Science Edition), Vol. 28 (2003) No 4, p.655-657

[2] Qian Aibing, New Technology of Library and Information Service, Vol. 24 (2003) No 2, p.42-44

[3] Liu Ziyu, Wang Qianling, Liang Puxuan, Application Research of Computers, Vol. 21 (2003) No 12, p.218-220

[4] Liu Chang, Zhang Meng, Journal of Jilin University (Information Science Edition), Vol. 31 (2013) No 3, p.320-323