An investigation and experiment of techniques for extracting characteristic words of documents

Hiroshi Iuchi (0710951)

School of Information Science,
Japan Advanced Institute of Science and Technology

November, 2011

Keywords: Characteristic word extraction, a law sentence, index charge account.

The modern world is overflowing with large volumes of information. The amount of information from documents being exchanged daily is increasing gradually with the increase in population and the development of communications methods and it is necessary to try to achieve effective information gathering.

When a user is gathering information, it is actually difficult to judge what is necessary among the large amount information available. Therefore, the user uses the information retrieval method to find information relating to the information required.

The documents obtained using information retrieval method are ranked and the user is able to obtain the desired information by looking through the results in order.

In order to reach the necessary information quickly using information retrieval method, there needs to be a guide for the information organized in advance and the required information needs to have already been gathered. However, it takes costs to create these and also it is difficult to know in advance whether it truly is the required information. If characteristic words can be extracted mechanically, this is highly useful for gathering information in advance and it is important to study and examine methods for extracting mechanical characteristic words.

The field of legislation is one of the fields where a large amount of information is overflowing. Electronic data and information retrieval systems for legal documents are available, but the indexes for each law are inadequate and currently the retrieval cannot be said to be sufficient. The resources for legal documents, such as characteristic word dictionaries and tagging, has also not been developed sufficiently.

Therefore, it can be considered extremely important to apply methods for extracting characteristic words to legal documents and evaluate them.

From the above viewpoint, this research aims to investigate previous researches on characteristic word extraction , apply the methods to actual legal documents and evaluate them.

There has been a lot of studies on methods for extracting characteristic words. We have classified them according to the purpose for extracting characteristic words.

(1) Information retrieval

Characteristic word extraction methods are used to characterize documents for the purpose of information retrieval. For this method, we investigated methods such as TF-IDF, a method using SVM, KeyGraph and a method using chi-square value.

(2) Information extraction

Characteristic word extraction methods are used to extract information for the purpose of extracting pre-determined information from documents. We investigated named entity extraction for this method.

(3) Terminology extraction

A method for extracting feature word are used to automatically extract terminologies from a specialized field corpus. We investigated a method based on appearance frequency and adjacency frequency.

According to the above investigation, considering that the chi-square value method does not require a large amount of statistical data relating of documents or corpus beforehand, we applied it to legal documents such as National Pension Act, Labor Standards Act and Building standards Act.

In the test results we could not obtain the level of accuracy with the chi-square value method used in this test that would enable it to be applied as an actual retrieval system for legal documents.

However, the method used in this test is not a method that can be replaced with such methods as TF-IDF, but it is considered that more suitable characteristic word extraction can be achieved by combining it with other methods.