Automatic assignment of HTML documents to web hierarchal category

KATAYAMA Kenichi

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

February 13,1998

Keywords: HTML, web, hierarchical categorization, tag, hyper text.

Abstract

Categorization of a document is to tidy a document along the classification that was given previously by catching the characteristic of a document. The trouble in time that finds out a certain intended document later by tidying a lot of documents is able to reduce with conceivable. A trouble is work that is required very to carry out the assignment to the category of a document with a help. It is expected that the classification that gave an objectivity and able to reduce the trouble in classification work by automating the assignment to this category becomes possible. However, must comprehend contents of a document, to carry out automation a high classification of accuracy is very difficult. Furthermore only a letter is not in the case of the document on web and be composed of the image, voice, program etc. and be making a problem furthermore difficult. Burying the key word that becomes the theme of a document by an image especially it is many it is said that the absence of information is awake with only a letter even that is crowded.

The classification of the layer of the document that is seen to directory searching of Web well in this research, is done automatically. Even condition where is able to classify only to assignment, a certain layer to plural node of the classification, document to the inside node and be not a classification of only a leaf node in this classification permits. It is conceivable that this classification be possible utilize to the automation of HTML document classification of a voluminous volume that at present is done with a help and be the classification that is used with yahoo, yahho etc. that are a representative directory searching system immediately.

The case that research that assigns a document was done bill a label undoes a teacher without requiring and of clustering, a mainstream was. However, a label needs to be swung to a category, to intend automatic classification of a directory searching system.

The classification to a layer is done automatically by using, the teacher existence study method that used the case that classified with a help previously and get a lot of cases by using the data of the directory searching system yahho that is carrying out classification of numerous documents and be the layer where it decided previously with a help in this research.

It was handling it targeting a book, news item, electronic mail, electron news in the research of many document automatic classification of established. Many automatic classification systems are classifying and are carrying out statistical processing by method of the characteristic vector and various resemblance degree calculation that have been utilized in the field of information searching. How I take out the characteristic certain key word from a document, in the case that I use the characteristic vector becomes a problem.

There is a characteristic analysis method of the text in the classification of WWW page of Ochitani, as the research of the classification that handles HTML document. He carries out resemblance degree calculation by using cosine distance and be using a morpheme, phrase, bigram as, the characteristic element. However, only the link information that is the hypertext structure that is one of the characteristic of WWW page is utilizing. Because the characteristic element takes out it to utilize the characteristic of WWW page more deeply in this research, I carry out characteristic extraction by a document restoration command called HTML document unique tag. There be the one etc. that emphasizing the one, a word that generate a title to, tag it thinks help that finds out the key word that shows the characteristic of a document. As for in web layer inclination is big to a data volume each of a node, even a distance scale other than a thought, cosine distance do constracting when this case, exerts an influence not a little on the performance of cosine distance and carry out the comparison. Also, a document is being assigned to the category of k piece to a high order of a resemblance degree in the method of Ochitaniy. It has assigned a category whenever a rank is high even if a resemblance degree is low when it is this method. There is condition where it assigns in the inside node where if the resemblance nature with a subordinate position node is not high in the case that it had classified it from an upper class node to a subordinate position node, as the characteristic of the layer that I use with this research does not go to a subordinate position over it. I am not able to carry out this arrangement method in the strategy that has assigned k piece without fail. I use the strategy that if there be not the one and assign, exceed to the category that exceeds the threthold and set up some sort of threthold to the ranking of a resemblance degree for this do not assign it. Furthermore, document unique classification rule on web is saved as knowledge in addition to, a statistical method and combine and utilize with a statistical method and think the efficacy.

It is used in the field of information searching to, an evaluation guideline and use reappearance rate (recall) and matching rate (precision).

Proposing the method of the characteristic extraction that uses tag with bill the weight of the word by a statistical method as the instrument of characteristic extraction of a document, in this paper, the efficacy is thought by using HTML document unique information in addition to a statistical method. One of the method of the information extraction, that added the information of tag to an appearance frequency even than

information extraction of only an appearance frequency of the word from formerly the improvement of accuracy was observed with recall, precision. Furthermore, by using threthold that differs every each category