International Journal of Innovative Computing, Information and Control
Volume **3**, Number **2**, April **2007**

ICIC International ©2007 ISSN 1349-4198
pp. 307–317

COMPARISON WITH FUZZY REASONING AND MODIFIED TF-IDF IN PAGE GROUPING FOR THE RESULT OF WEB RETRIEVAL

Tsutomu Miyoshi and Hiroo Joichi

Department of Information and Knowledge Engineering
Tottori University
4-101, Koyama-cho Minami, Tottori-shi, Tottori 680-8552, Japan
mijosxi@ike.tottori-u.ac.jp

Received May 2006; revised August 2006

Abstract. *For Web page retrieval, even if the user uses the same keywords, different kinds of pages tend to be mixed in retrieval result because of polysemy or ambiguity of words. We already proposed the system of having improved the problem, which classifies retrieval result to groups automatically according to page contents using the vector space model method, the frequency of word appearance and the fuzzy reasoning. In our system, to reduce dimension of vector space, index words are selected based on the importance of a word measuring by fuzzy reasoning. On the other hand, TF-IDF is often used as a measure of the importance of a word in Information Retrieval. Then we thought that it was necessary to compare with the performances of TF-IDF and fuzzy reasoning. From the experiments, we confirmed that similar classification for retrieval pages in terms of human sense by the system with fuzzy reasoning rather than with modified FT-IDF.*

Keywords: Web page retrieval, Fuzzy reasoning, TF-IDF, Extract a feature

1. Introduction.

Search engines are mainly used for Web page retrieval in recent years. For the present situation which the amount of homepage increases so quickly, the problems are pointed out [1] that required Web pages are not displayed on a higher rank in retrieval result. One of the reasons is that, the pages in retrieval result are selected because only they include searching keywords in them. Some techniques to solve this problem were reported, for example, vector space model method [4] using the frequency of word appearance [5-7], information gathering or retrieval using fuzzy linguistic representation [9-11], clustering by multi features [14], classification by hyperlink [15], etc.

To solve the problem we paid our attention to the polysemy or ambiguity of searching keyword. Even if the user uses the same keywords, different kinds of pages tend to be mixed in retrieval result because of polysemy or ambiguity of words. We thought that, the appearance frequency of specific words, which represent the contents of the page, become high for the pages of similar content. So, classification of pages is possible by the co-occurrence frequency of words appearance using the vector space model method.

The vector space model method, as conventional method, is classifying pages by creating the feature vector of each page based on the frequency of word appearance, and measuring the degree of similarity with other pages by feature vectors. However this method has two problems to apply page classification. One is that, computational cost of calculating the similarity is too high, because all words appearing to all pages are used as the vector