International Journal of Innovative Computing, Information and Control
Volume **2**, Number **1**, February **2006**

ICIC International ©2006 ISSN 1349-4198
pp. 237–247

THE IMPROVEMENT OF WEB PAGE RETRIEVAL BY PAGE GROUPING USING FUZZY REASONING

Tsutomu Miyoshi and Hiroo Joichi

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

Received February 2005; revised September 2005

Abstract. *For Web page retrieval, we consider that, even if the user uses the same keywords, different kinds of pages tend to be mixed in retrieval result because of the polysemy or ambiguity of words. In this paper, we propose the system which classifies retrieval results to the group according to page content by using the vector space model method, the frequency of word appearance and the fuzzy reasoning. From the experiments, we confirmed that similar classification for retrieval pages in terms of human sense by the proposed system.*

Keywords: Web page retrieval, Fuzzy reasoning, Frequency of word appearance, Extract a feature

1. **Introduction.** Search engines have been mainly used for Web page retrieval in recent years. For the present situation where the amount of homepages increase so quickly, the problems are pointed out [1] that required Web pages are not displayed at a higher rank in retrieval result. One of the reasons is that, a retrieval result is selected only by the reason that it includes searching key words 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], etc. We tried to solve the problem by the vector space model method that is suitable for classifying many and unspecified documents.

We paid attention to the polysemy or ambiguity of searching keywords. Even if the user uses the same keywords, different kinds of pages tend to be mixed in retrieval because of polysemy or ambiguity of words. We considered that, as for the page of similar content, the appearance frequency of specific words, representing the content of the page, become high. We thought that classification of pages is possible by the co-occurrence frequency of word appearances.

The vector space model method, as a 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. One is that, the cost of calculating the similarity is too high because all words appearing in all documents are used as the vector space. A reduction in calculation costs should be considered because a quick response is better for Web