International Journal of Innovative Computing, Information and Control
Volume **6**, Number **6**, June **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 2749–2760

WEB PAGES CLUSTER BASED ON THE RELATIONS OF MAPPING KEYWORDS TO ONTOLOGY CONCEPT HIERARCHY

Rung-Ching Chen[1], Cho-Tsan Bau[1], Ming-Yung Tsai[1]
and Chung-Yi Huang[2]

[1]Department of Information Management
Chaoyang University of Technology
168 Jifong E. Rd., Wufong Township Taichung County 413, Taiwan
crching@mail.cyut.edu.tw

[2]Department of Computer Science and Engineering
National Chung-Hsing University
250 Kuo Kuang Rd., Taichung 402, Taiwan
chungyi@ctu.edu.tw

Received January 2009; revised July 2009

Abstract. *In this paper, we present a related web page cluster method that not only considers the corresponding domain keywords on ontology but also analyzes semantic contents of web pages. First, the method embeds the corresponding domain ontology of search keyword to find web pages from the Internet. Next, consider the location of the keywords in the web pages, and relations between keywords and concepts in the domain ontology to find the features of the web pages. Then, the web pages were clustered based on the similarity values of mapping keywords concept-ontology level relations. Primary experimental results prove that our method is effective to find related web pages.*

Keywords: Related web pages, Ontology, Semantic search, RDF

1. **Introduction.** Both the information and the number of users on the Internet are rapidly growing. Hence, using a search engine to find related web pages and retrieve useful information has become important task [1]. To understand the mind of users and to effectively find desired web pages are an enormous challenge [2,3]. Since domain-specific knowledge is not embedded in the search engine, keyword search return too broad selections including unrelated information. This happens because search engines cannot understand the latent semantic of keywords and user's intention. For example, the word "wind" is a phenomenon in the meteorology domain, but a type of instrument in the music domain. So, not only keyword but domain information is important for optimizing retrieval.

Before return searching results, web pages cluster is needed. Traditional web page search methodology is based on keyword, anchor text or hyperlink information [4]. Researchers are currently evolving intelligent search methods to provide a search engine that understands context of search keywords. Oyama has proposed a method that uses a set of training web pages collected in advance by a human [5]. In the training phase, important and unimportant keywords are extracted from the domain. Then the keywords are combined using a Boolean function to form keyword spices and to generate a decision tree. The keyword spices and the decision tree are then sent to the search engine to find related web pages. Khan and Lee collected a list of high order search results and used the frequency of keywords as an important characteristic from those results [6]. They input the listed of keywords and web pages to a neural network to find related web pages.