4137578 SCCS/M : MAJOR : COMPUTER SCIENCE ; M.Sc.(COMPUTER SCIENCE)
KEY WORDS : METASEARCH ENGINE / CLUSTERING / INFORMATION RETRIEVAL
TANASAI SUCONTPHUNT : CONTENT-BASED CLUSTERING METASEARCH ENGINE. THESIS ADVISORS: THANWADEE SUNETNUNTA, Ph.D., DAMRAS WONGSAWANG, Ph.D. 120 P. ISBN 974-665-025-4

Using a search engine as a tool to find information on Internet is the most favorite way for Internet users. However, using only a single search engine is not sufficient since no single search engine can cover an entire set of Web pages on Internet. Moreover, to query to every search engine is a time consuming process. In addition, each search engine has its own ranking algorithm and specific searching area. Thus, the results of each search engine are differrent in ranking score and presentation format, thereby making the users confused easily.

Metasearch engines have been developed to resolve the disadvantages of search engines by enhancing time saving, searching coverage, and uniform result presentation. These features help users gain more coverage and a variety of returned results, thereby increasing the possibility to find relevant information more than using a single search engine. Moreover, the results are presented in one uniform format.

However, most metasearch engines present their results by using a ranked list. Since the results of metasearch engines are always diverse by mixed results, the ranked list forces users to spend time to sift through a diversity of the results and make it hard for the users to find the information they are looking for.

Clustering technique is a way to alleviate this problem. It divides the results into different labeled groups, i.e. meaningful clusters, which will help users find relevant documents easier and faster. Nevertheless, there is no single traditional clustering algorithm available today which truly creates sufficient easy-to-read cluster labels for users.

In this thesis, we propose a Content-Based Clustering (CBC) technique which aims to yield clusters that help users find a relevant document easier and faster by applying (i) a feature extraction tool, (ii) a hybrid clustering of nonhierarchical clustering method and a single link HAC. In addition, we develop a prototype of an off-line metasearch engine, namely Content-Based MetaSearch Engine (CBCMSE), as a proof of our concept.