Abstract

Document clustering is a technique that can be implemented to ease user in finding the expected documents from many documents retrieved by the search engine. This technique groups documents based on certain category, so that the user's searching on retrieved documents can be simplified.

Semantic Hierarchical Online Clustering (SHOC) algorithm is one of clustering algorithms which groups web documents into certain cluster based on key phrases contained in the documents. This research implements and analyzes web search result clustering using SHOC algorithm.

The testing result shows that SHOC algorithm is able to separate the relevant and unrelevant retrieved documents where the performance depends on the quality of the search engine and the documents' characteristic. SHOC algorithm is suitable for grouping documents that share key phrases each other. And for the influence of the quality of search engine, poor search engine's precision will generate many "waste clusters", and poor search engine's recall will decrease the accuracy of the generated clusters. To handle the poorness of the search engine, the value of cluster quality threshold in SHOC algorithm needs to be set according to the quality of search engine, so the relevant documents can still be grouped.

***Keywords**: search engine, retrieved documents, clustering, key phrases, SHOC algorithm*