Abstract

Document clustering is an appropriate way to simplify the search engine performing the query against a large collection of documents. Similar documents will be grouped to form different topics or subtopics. Document clustering algorithms that are often studied are the batch clustering ones, where the entire document is required from the beginning and the clustering is performed by many iterations of each document. However, with the current groeing online publishing on the web, explosion of information is increasing every day. Batch clustering methods are considered inefficient for such cases. In order for the clustering process can be performed immediately after the document signed in, it needs to be done incrementally.

There are several popular incremental clustering algorithms. One of them is the Cobweb algorithm implemented in this final project. Cobweb is used to classify retrieved documents from search results by search engine. Cobweb perform data clustering by building a classification tree where every node of the tree depicts the cluster that contains the data objects. In the tree building, Cobweb uses category utility (CU) to evaluate the tree and get the most appropriate grouping of data. From the testing performed on this final project, final result shows that the Cobweb clustering algorithm implemented on retrieved documents provides solutions with good quality, because despite the inevitable overlapping clusters of trees, still proved to have the cohesiveness characteristic. Cohesive is a state where the similarity between documents in the same cluster is greater than the similarity between documents in different clusters.

***Keywords** : document, incremental clustering, Cobweb, search engine, retrieved documents ,classification tree, category utility, similarity, cohesive*