UMS

UNIVERSITI MALAYSIA SABAH

ENHANCING KNOWLEDGE MANAGEMENT BY DEVELOPING AN AUTOMATED DOCUMENT LABELLING BASED ON CONCEPTS AGGREGATION USING HAC TECHNIQUE

FINAL REPORT

GRANT NO: SLB0008-TK-1/2011

RAYNER ALFRED, LEAU YU BENG AND TAN SOO FUN

SYNOPSIS

Vast amounts of text documents are available in various fields. The accumulations of available text documents have raised new challenges for information retrieval (IR) technology. Therefore, in order to facilitate the knowledge management process, various approaches and techniques applied on text classification (categorization) and text clustering are being compared and studied. The most common way to organize and label documents is to group similar documents into clusters by clustering them and then extract concepts that characterize each cluster. Normally, the assumed number of clusters may be unreliable since the nature of the grouping structures among the data is unknown before processing and thus the partitioning methods would not predict the structures of the data very well. Hierarchical clustering has been chosen to solve this problem by which they provide data-views at different levels of abstraction, making them ideal for people to visualize the concepts generated and interactively explore large document collections. Another problem that needs to be considered is the appropriate method of combining two different clusters to form a single cluster. In order to perform this task, various distance methods will be studied in order to cluster documents by using the hierarchical agglomerative clustering. Clusters very often include sub-clusters, and the hierarchical structure is indeed a natural constraint on the underlying application domain. In order to manage and organize documents effectively, similar documents will be merged to form clusters. Each document is represented by one or more concepts. The goal of this project is to generate concepts that characterize English documents by using the hierarchical agglomerative clustering. One of the advantages of using hierarchical clustering is that the overlapping clusters can be formed and concepts can be generated based on the contents of each cluster. Besides that, different distance measures will be used in order to investigate the quality of clusters produced.