UMS

UNIVERSITI MALAYSIA SABAH

DEVELOPMENT OF A TEXT ANALYZER FOR AUTOMATIC CATEGORIZATION OF TEXTS DOCUMENTS BASED ON INTERACTIVE VISUALIZATION APPROACH

FINAL REPORT

GRANT NO: SBK0012-SG-1/2011

By

MOHD NORHISHAM BIN RAZALI @ GHAZALI, RAYNER ALFRED, SURAYA ALIAS AND ASNI TAHIR

SYNOPSIS

The wide availability of huge collections of text documents (news corpora, e-mails, web pages, scientific articles and etc) has fostered the need for efficient text mining tools. Information retrieval, text filtering and classification, and information extraction technologies are rapidly becoming key components of modern information processing systems, helping end-users to select, visualize and shape their informational environment. The ability to visualize documents into clusters is very essential. The best data summarization technique could be used to summarize data but a poor representation or visualization of it will be totally misleading. As proposed in many researches, clustering techniques are applied and the results are produced when documents are grouped in clusters. However, in some cases, user may want to know the relationship that exists between clusters. In order to illustrate relationships that exist between clusters, a hierarchical agglomerative clustering technique can be applied to build the dendogram. This dendogram display the relationship between a cluster and its sub-clusters. For this reason, user will be able to view the relationship that exists between clusters. In addition to that, the terms or features that characterize each cluster can also be displayed to assist user in understanding the contents of whole text documents that stored in the database. In this research, a Text Analyzer (VisualText) that automates the categorization of text documents based on a visualization approach using the Hierarchical Agglomerative Clustering technique will be proposed. With VisualText, users are able to analyze and categorize text documents automatically, visualize the overall structure of their informational environment by visualizing each cluster and its sub-clusters, identify words or terms used to categorize each cluster and its sub-cluster and finally evaluate the quality of the text categorization based on the distance method. The proposed tool is potentially very useful for analyzing text documents automatically for summarization purposes and thus facilitates decision making process.