A Visualization Approach to Automatic Text Documents Categorization Based on HAC

Abstrak

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 (HAC) technique can be applied to build the dendrogram. The dendrogram produced 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 paper, a Text Analyzer (VisualText) that automates the categorization of text documents based on a visualization approach using the Hierarchical Agglomerative Clustering technique is proposed. This paper also studies the effect of using different inter-cluster proximities on the quality of clusters produced. Cophenetic Correlation Coefficient is measured in order to evaluate the quality of clusters produced using these three different inter-cluster distance measurements.