ABSTRACT

Rapid growth in IT (Information Technology) affects the amount of computerized data and digital documents. To obtain the desired information from those data is certainly not an easy task. Clustering is a technique in Text Mining that can be a solution to overcome this problem. Clustering aims to classify data/document into a group/cluster that has similar information/characteristics. The effectiveness and results of the clustering process is determined by the applied algorithm, one of the popular algorithms for document clustering is Suffix Tree Clustering (STC). However, the actual of STC algorithm itself still has some shortcomings, one of those shortcomings is STC does not have an efficient similarity measurements to compute the similarity between the inter-clusters and intra-clusters. Thus, this study will use a modified algorithm of STC which is Suffix Tree Hierarchical Agglomerative Clustering (STHAC). This algorithm will add two additional processes on a regular STC algorithm, which are Clusters Ranking and Filtering process and Cluster Cleaning process. Results of clustering were analyzed by test validation using Precision, Recall and F-measure to determine the quality of its performance. After conducting some tests, it showed that in general STHAC algorithm had better performance compared to STC algorithm. The value of Recall and F-measure on the test showed that STHAC algorithm scored higher than STC. As for the Precision value, STHAC algorithm scored lower than STC.

Keyword: *Clustering, Text mining, Suffix Tree Hierarchical Agglomerative Clustering, Precision, Recall, F-measure*