Abstract

Data mining is an exploration process and can be applied at huge data analysis to gain useful pattern. Data mining is a combination process between several fields for example machine learning, statistical analysis, and database.

One of the important task in data mining is clustering. Clustering is a process to partition data objects into clusters. The similar objects will be placed into the same cluster and the different objects will be placed into different clusters.

The aim of this final project is grouping documents by Top-k scoring algorithm. The documents here are taken from work security documents of PT.Pertamina UP IV Cilacap, because it contents of unstructured complicated text documents which needs big effort to find out solution documents manually to face a problem. Similarity or distance is measured by simple additive count of words found in both documents that are compared.

After the testing with several threshold, the results shows that top-k scoring algorithm can be used to group Indonesian documents with accuracy up to 96.67%. The calculation of the accuracy was done by comparing the clustering result with the result of manual grouping.

Key words *: data mining, clustering, top-k scoring, cluster, WIDF.*