Abstract

Large amount of data especially Indonesian news documents is a certain reason why clustering technique become more popular. With this technique, Indonesian news documents will be easier to be grouped although the class label is unknown. There are many clustering methods that can be used, but usually these methods are do not handle yet the high dimension of data, non-meaningful cluster description, and still allow overlap (the condition where one document may in to some groups).

These problems can be handled using *Frequent Itemset-Based Hierarchical Clustering* (FIHC). High dimension of data and non-meaningful cluster description can be handled by eliminating non-frequent words, and overlap condition can be handle by disjoint cluster.

The FIHC's output is visualized by hierarchy tree. Based on experiment, cluster from FIHC has a good quality, especially if it is compared with another algorithm, *Hierarchical Frequent Term-Based Clustering* (HFTC). The meaningful cluster description is produced, and overlap is definitely none. Larger amount of dataset on experiment, the minimum support value will be decrease. And for the same value of minimum support, if cluster support decrease, F-Measure also decrease.

Keywords: *clustering, frequent term-based text clustering, FIHC, F-Measure.*