Abstract

An increasing number of documents in text format significantly lately makes the process of grouping documents (*document clustering*) becomes important. Grouping the document aims to divide the document into several groups (clusters) so that the documents possessed a high degree of similarity are included in the same cluster and possessed similarities that have low included indifferent clusters. To perform such *clustering*,*clustering* algorithms used one of the *CanopyClustering*. *Canopy Clustering* is a development of the *K*-means *clustering*. This algorithm can overcome the problems found on *the K*-means in amatter of accuracy and processing time for large data sets. *Clustering* of the value of the parameter T.This parameter serves as the cluster size on the formation of Canopy. To measure the similarity between the documents before the clustering process used *Euclidean distance*.

In this final cluster resulting accuracy is measured using *precision*,*recall*, and *F1*-measure. Based on experiments conducted that *Canopy Clustering* using *K*-means higher level of accuracy andless time to process compared to the *K*-means algorithm .

Keywords: *Canopy Clustering*, *K*-means, *Clustering*