Abstract

Clustering is a mean of process to collect different objects based on information extracted from the data. In a set of data consists of several news documents, clustering process can be used to collect some similar news together in order to help user pick some related news based on his preferences. However, as the data that is being used becoming more complex, problems in finding good clusters are arising. A difficulty to find noise documents or outliers is a one common problem. In order to overcome this problem, a method based on shared nearest neighbor approach will be introduced. By using this method, clusters are formed automatically based on documents linkage with their neighbors with regard to its three parameters (i.e. K, EPS, and MinPts). Using a dataset taken from www.reuters.com and SMART, the test results, that is determined by comparing cohesion and separation values, show that the best result is achieved when the parameters are similar or almost similar. Alas, in general, there are no fixed values for these parameters that will yield the best result.

Keywords: clustering, shared nearest neighbor, noise, parameter, cohesion, separation.