Abstract

With the development of technology, large numbers of information like news articles are available on the internet. Hence text categorization is needed by applying classification as one of data mining task. However, the major problem of text categorization is the high dimensionality of data. Therefore, we need to select some representative attributes to improve performance of text categorization. One of technique to do this is feature selection.

Feature selection can reduce the high dimensionality, so, the effectiveness of classifier improves. A number of Feature selection mechanisms have been explored in text classification, among which mutual information, information gain, and chi-square are considered most effective.

In this final assessment, will be discussed the other methods known by the name of Within Class Popularity that is used for text classification. The dataset used is Indonesian language news articles from the web and using multiple classifiers, such as Naive Bayes, and k-Nearest Neighbor (kNN).

In addition, also conducted a comparative analysis of within class popularity methods with mutual information and chi-square by using the classifiers and based on the analysis results is known that within class popularity have better performance than the mutual information and chi-square.

Keywords: Text classification, within class popularity, Feature selection, classifier