Abstract

One of the significant problem in the text categorization is high dimentionality of data that cause a long processing time. One of the severals ways to overcome this problem is doing feature selection phase to the data before categorization process. The goal of feature selection is to produce important and relevant features. Therefore, the data dimensionality can be reduced.

In this final task, the research is about feature selection using Information Gain and Chi-Square in N-Gram text categorization. Categorization is done by counted the distance of category profile and the document profile, where the profiles is made from the features existed. Therefore, the number of the feature is have a high influence in the time needed for categorization process. In the text categorization using N-gram without feature selection, the result shows that F-measure give a value of 0.89, where 2-gram is used here. When feature selection is done by Information Gain to the number of 80 %, F-measure value increase up to 0.935. And When feature selection is done by Chi-Square to the number of 20 %, F-measure value increase up to 0.94.

Selecting feature using Information Gain feature selection is faster than selecting feature using Chi-Square feature selection. However, the performace of text categorization using features from the feature selection proses by Chi-square is better.

Keywords: feature selection, 2-gram, Information Gain, Chi-Square, F-Measure.