Abstract

In the Text Preprocessing, term weighting is a step that is very important. This step is applied in order to give a value/weight on the term that is contained in a document. The weight given to a term depends on the method that is used for the weighting. In the text mining, there are some term weighting method such as TF, TF·IDF, and WIDF.

In this Final Task, some term weighting methods like TF, TF·IDF, and WIDF, are compared each other by seeing the output of the text categorization performance. Some parameters that will be used as a measurement for comparing the text categorization performance are recall, precision, and f-measure. To test the output of the weighting result, it will be used a classification tool like Weka, with NaiveBayes and Naïve Bayes Updateable as its classifier.

Based on the result, it is concluded that the WIDF weighting method has better performance compared with others weighting method (TF and TF·IDF). Usually, WIDF surpass other methods in almost every testing phase. The benefit of WIDF that counting the term presence frequency in a document and normalize it over document collection, become the advantage compared with other methods. So, this method is better than the others.

Keyword: *text preprocessing, term, term weighting, TF, TF·IDF, WIDF*