Abstract

Text categorization (or also known as text classification) is a task that sort a collection of documents D into a category that has been determined automatically Φ : D x C. In text categorization, one of the process are text preprocessing which include feature selection and term weighting.

One known method of term weighting is TF-IDF, in this method where each term / word in a document is calculated in their frequency in a document (term frequency), which the results combined with the occurrence frequency of terms in a document collection (inverse document frequency). Term that often appears on the document but rarely appear on the set of documents providing a high weight value. TF-IDF will increase with the number of occurrences of terms in a document and is reduced by the number of terms that appear on the document collection.

Since text categorization are supervised which include dataset was divided into training and testing datasets, we need a method that meets the requirement. In standard IR contexts this assumption is reasonable, since it encodes the quite plausible intuition that a term *tk* that occurs in too many documents is not a good discriminator, i.e when it occurs in a query *q* it is not sufficiently helpful in discriminating the documents relevant to *q* from the irrelevant. However, if training data for the query were available (i.e. documents whose relevance or irrelevance to *q* is known), an even stronger intuition should be brought to bear, i.e. the one according to which the best discriminators are the terms that are distributed most differently in the sets of positive and negative training examples. Training data is not available for queries in standard IR contexts, but is usually available for categories in TC contexts, where the notion of "relevance to a query" is replaced by the notion of "membership in a category". In these contexts, *category-based* functions like on *Term Evalution Function* such as *Chi-square, Information Gain (IG),* dan *Gain Ratio (GR)* that score terms according to how differently they are distributed in the sets of positive and negative training examples, are thus better substitutes of *idf*-like functions.

In this method combined the term evaluation function value of each term is chosen with a value frequencynya term in each document and then performed SVM classification methods and evaluated with the parameters of precison, recall, f-measure and accuracy. The results showed that the supervised term weighting method gives better performance than TF-IDF, especially at the local threshold policy.

Keywords : text categorization, term frequency, inverse document frequency, term weighting, supervised term weighting.