ABSTRACT

Classification of text documents is a simple but very important problem because the benefits are quite large considering the number of documents that exist every day is increasing. However, most existing document classification techniques require large amounts of labeled documents to perform training and testing phases. In doing classification of documents, in this final project used Principal Component Analysis algorithm combined with Naïve Bayes for supervised document. Principal Component Analysis is a technique that can be used to extract the structure of a high-dimensional data without losing any significant information to the whole data and then an algorithm that can produce the prediction and accuracy of the document is Naïve Bayes. The Naive Bayesian classification is a simpler classification based on the Bayes theorem application with the assumption between independent variables. In this case, it is assumed that the presence or absence of a particular event of a group is not related to the presence or absence of any other event. The results of system testing yielding data reduced by Principal Component Analysis (PCA) have a same accuracy for certain datasets than without PCA.

. The data that used is R8 data from Reuters-21578 Text Categorization of Collection Data Set. The best accuracy in this research was made by Naïve Bayes method with combination no 1 59,111%, while for Naïve Bayes + PCA method which combination no 10 with 64 %.

Keywords :
Document Classification, Principal Component Analysis, Naive bayes