DOCUMENT CLASSIFICATION USING BACKGROUND SMOOTHING

Andi Pramurjadi, Julio Adisantoso

Department of Computer Science

Faculty of Mathematics and Natural Sciences

Bogor Agricultural University

julioipb@gmail.com

2009

Abstract

Naïve Bayes Classifier (NBC) is one of the methods for text or document classification. A common problem that often occurs on NBC method is data sparsity, especially when the size of training data is too small. One way to handle the sparsity problem is to use background smoothing technique. The aims of this research are to look at the background smoothing effect on short and long query, and to compare it with NBC on small training data.

In this research, we use documents from the Agricultural Research Journal of horticultural domain. The results indicate that the accuracy of document classification on NBC+Background Smoothing is 92.3%, not significantly different from that obtained using only NBC. Improvement of the accuracy is only 1.78% from the results obtained on NBC. However, the results of the classification with NBC+Background Smoothing has been able to properly classify documents of Agriculture Research Journal at horticultural domain, so that it can be used to organize documents much easier for users to find information related to the documents.