# Recomputation of Class Relevance Scores for Improving Text Classification

Sang-Bum Kim and Hae-Chang Rim

Dept. of Computer Science and Engineering, Korea University,
Anam-dong 5 ka, SungPuk-gu, SEOUL, 136-701, KOREA
{sbkim,rim}@nlp.korea.ac.kr

**Abstract.** In the text classification task, bag-of-word representation causes a critical problem when the prediction powers for a few words are estimated terribly inaccurately because of the lack of the training documents. In this paper, we propose recomputation of class relenvace scores based on the similarities among the classes for improving text classification. Through the experiments using two different baseline classifiers and two different test data, we prove that our proposed method consistently outperforms the traditional text classification strategy.

## 1 Introduction

Text categorization (or classification, filtering, routing, etc.) is the problem of assigning predefined categories to free text documents. This problem is of great practical importance given the massive volume of online texts available through the World Wide Web, internet news feeds, electronic mail, corporate databases, medical patient records, digital libraries, etc. Learning methods are frequently employed to automatically construct classifiers from labeled documents. A growing number of statistical learning methods have been applied to this problem in recent years, including nearest neighbor classifiers[4], perceptron classifiers[3], Bayesian probabilistic classifiers[2], and support vector machines[1], etc. Most text classification approaches are based on bag-of-word representation of documents. Each word has its own class prediction power learned by specified learning algorithm, and they are combined to predict the class of each document. A bag-of-word representation scheme is widely used in many other IR applications because of its simplicity and efficiency. However, this representation causes a critical problem when the prediction powers for a few words are estimated terribly inaccurately because of the lack of the training documents. For example, suppose that the word "*model*" has appeared all the three training documents for class `airplane`, but has not appeared in any other documents. In this case, the word "*model*" has extreme weight, i.e., prediction power, to favor the class `airplane`. Thus, if a new test document including the word "*model*" will be classified into the class `airplane` although the test document is about "*Recent various computer models*". It is obvious that the proper class can be assigned to the test document if the document has many informative terms such as "*computer*", "*CPU*" which favor the class `computer`. For this problem, some studies