Raising High-Degree Overlapped Character Bigrams into Trigrams for Dimensionality Reduction in Chinese Text Categorization

Dejun Xue, Maosong Sun

National Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University
Beijing, China 100084
xdj00@mails.tsinghua.edu.cn; lkc-dcs@mail.tsinghua.edu.cn

Abstract. High dimensionality of feature space is a crucial obstacle for Automated Text Categorization. According to the characteristics of Chinese character N-grams, this paper reveals that there exists a kind of redundancy arising from feature overlapping. Focusing on Chinese character bigrams, the paper puts forward a concept of δ -overlapping between two bigrams, and proposes a new method of dimensionality reduction, called δ -Overlapped Raising ($\delta - OR$), by raising the δ -overlapped bigrams into their corresponding trigrams. Moreover, the paper designs a two-stage dimensionality reduction strategy for Chinese bigrams by integrating a filtering method based on *Chi-CIG* score function and the $\delta - OR$ method. Experimental results on a large-scale Chinese document collection indicate that, on the basis of the first stage of reduction processing, $\delta - OR$ at the second stage can significantly reduce the dimension of feature space without sacrificing categorization effectiveness. We believe that the above methodology would be language-independent.

1 Introduction

Nowadays, a large volume of information in digital form is available. Automated Text Categorization (*TC*), which automatically classifies natural language documents into a predefined set of thematic categories, is becoming a key approach in content-based document management tasks [1]. A range of statistic and machine learning methods have been employed in *TC*, such as Probabilistic Model [2, 3], Neural Network [4], K-Nearest Neighbors [5], Decision Tree [6], Centroid-Based Classifier [7], Rocchio Classifier [8], Support Vector Machine [9], as well as Classifier Committee [10], etc. In this kind of inductive classifier learning, text documents are usually indexed to weighted feature vectors with an indexing language, such as words or phrases [11]. Although phrases have intuitively better semantic qualities than words, previous works show that more sophisticated representations have not led to significant improvement over word indexing due to their inferior statistical qualities on some occasions [1, 12].