Enhancement of DTP Feature Selection Method for Text Categorization[⋆]

Edgar Moyotl-Hernández, Héctor Jiménez-Salazar

Facultad de Ciencias de la Computación,
B. Universidad Autónoma de Puebla,
emoyotl@mail.cs.buap.mx, hjimenez@fcfm.buap.mx

Abstract. This paper studies the structure of vectors obtained by using term selection methods in high-dimensional text collection. We found that the *distance to transition point* (DTP) method omits commonly occurring terms, which are poor discriminators between documents, but which convey important information about a collection. Experimental results obtained on the Reuters-21578 collection with the k-NN classifier show that feature selection by DTP combined with common terms outperforms slightly simple *document frequency.*

1 Introduction

The goal of *text categorization* (TC) is to classify documents into a set of predefined categories. In TC each document is usually represented as a vector of terms in a multidimensional space, in which each dimension in the space corresponds to a term. Typically even a moderately sized collection of text has tens or hundreds of thousands of terms. Hence, the document vectors are high-dimensional. However, most documents contain fewer terms, 1-5% or less, in comparison to the total number of terms in the entire text collection. Thus, the document vectors are *sparse* [3].

For reasons of both efficiency and efficacy, *feature selection* (FS) techniques are used when applying machine learning algorithms to text classification. In our previous experiments [6] we found that FS using DTP achieves performance superior to *document frequency*, and comparable to *information gain* and *chi-statistic*; three well known and effective FS techniques [10]. However, the vectors produced by DTP have a "sparse" behavior that is not commonly found in low-dimensional text collections.

In this paper, our first focus is to study the structure of the vectors produced by term selection methods when applied to large document collections. Such structural insight is a key step towards our second focus, which is to explore the relationships between DTP and the problem of the sparseness. We hypothesized that supplementing it with high frequency terms would improve term selection by adding important (and also common) terms; and we report experimental results

[⋆] This work was supported by VIEP-BUAP, grant III9-04/ING/G.