Full paper to appear in Proc. of the 27th European Conference on Information Retrieval. (ECIR 2005)

On Compression-Based Text Classification

Yuval Marton[1], Ning Wu[2], and Lisa Hellerstein[2]

[1] University of Maryland, Department of Linguistics, 1401 Marie Mount Hall, College Park, MD 20742-7505 ymarton@umiacs.umd.edu
[2] Polytechnic University, Department of Computer and Information Science, 5 Metrotech Center, Brooklyn, NY, 11201 wning@cis.poly.edu hstein@cis.poly.edu

Abstract

Compression-based text classification methods are easy to apply, requiring virtually no preprocessing of the data. Most such methods are character-based, and thus have the potential to automatically capture non-word features of a document, such as punctuation, word-stems, and features spanning more than one word. However, compression-based classification methods have drawbacks (such as slow running time), and not all such methods are equally effective. We present the results of a number of experiments designed to evaluate the effectiveness and behavior of different compression-based text classification methods on English text. Among our experiments are some specifically designed to test whether the ability to capture non-word (including super-word) features causes character-based text compression methods to achieve more accurate classification.