A Comprehensive Study of Features and Algorithms for URL-Based Topic Classification

EDA BAYKAN, Izmir University
MONIKA HENZINGER, University of Vienna
LUDMILA MARIAN, CERN
INGMAR WEBER, Yahoo! Research

Given *only* the URL of a Web page, can we identify its topic? We study this problem in detail by exploring a large number of different feature sets and algorithms on several datasets. We also show that the inherent overlap between topics and the sparsity of the information in URLs makes this a very challenging problem. Web page classification without a page's content is desirable when the content is not available at all, when a classification is needed before obtaining the content, or when classification speed is of utmost importance. For our experiments we used five different corpora comprising a total of about 3 million (URL, classification) pairs. We evaluated several techniques for feature generation and classification algorithms. The individual binary classifiers were then combined via boosting into metabinary classifiers. We achieve typical F-measure values between 80 and 85, and a typical precision of around 86. The precision can be pushed further over 90 while maintaining a typical level of recall between 30 and 40.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms: Experimentation

Additional Key Words and Phrases: Topic classification, URL, ODP

ACM Reference Format:
Baykan, E., Henzinger, M., Marian, L., and Weber, I. 2011. A comprehensive study of features and algorithms for URL-based topic classification. ACM Trans. Web 5, 3, Article 15 (July 2011), 29 pages.
DOI = 10.1145/1993053.1993057 http://doi.acm.org/10.1145/1993053.1993057

1. INTRODUCTION

Topic classification of Web pages is normally performed based on the *content* of the pages, with additional clues coming from the link structure of the Web graph [Chakrabarti et al. 1998; Qi and Davison 2006]. However, there are several advantages to attempt the classification task using only URLs, and this is the problem studied in this article.

One advantage of such an approach is speed. The length of a URL is a tiny fraction of the typical length of a Web page. This enables a much faster construction of feature vectors and also speeds up the classification itself, due to the reduced number of nonzero features. But there are also scenarios where the content of a Web page

---

This is an extended version of a poster published at WWW2009 [Baykan et al. 2009].
Authors' addresses: E. Baykan, Izmir University, Izmir, Turkey; M. Henzinger, Department of Computer Science, University of Vienna, Austria; L. Marian, CERN, Geneva, Switzerland; I. Weber (corresponding author), Yahoo! Research, Barcelona, Spain; email: ingmar@yahoo-inc.com.

DOI 10.1145/1993053.1993057 http://doi.acm.org/10.1145/1993053.1993057