Microsoft Silicon Valley Web Spam Challenge Entry

Steve Chien
Microsoft Research
1065 La Avenida
Mountain View, CA, USA
schien@microsoft.com

Dennis Fetterly
Microsoft Research
1065 La Avenida
Mountain View, CA, USA
fetterly@microsoft.com

Mark Manasse
Microsoft Research
1065 La Avenida
Mountain View, CA, USA
manasse@microsoft.com

Marc Najork
Microsoft Research
1065 La Avenida
Mountain View, CA, USA
najork@microsoft.com

Alexandros Ntoulas
Microsoft Live Search Labs
1065 La Avenida
Mountain View, CA, USA
antoulas@microsoft.com

Categories and Subject Descriptors

H.5.4 [**Information Interfaces and Presentation**]: Hypertext/ Hypermedia; K.4.m [**Computers and Society**]: Miscellaneous; H.4.m [**Information Systems**]: Miscellaneous

General Terms

Measurement, Experimentation

Keywords

Web characterization, web spam

1. ABSTRACT

This paper describes our contribution to the 2007 Web Spam Challenge. We computed some additional features from the data provided with the UK 2006-05 dataset, and other features from external data sources.

Our contributions to the Web Spam Challenge fall into two categories. First, we used the features introduced in our earlier work ([2], [3]). Second, we incorporated features that are economic in nature, namely domain registrar information and Google AdSense publisher ID.

In addition to the features provided by the organizers of the challenge ([1], [4]), we added 84 additional features. These features fall into the following groups:

- Features derived from re-crawling the 77 million URLs in the UK 2006-05 dataset.
- Features derived from the registrar records for the 7,707 domains in the UK 2006-05 dataset.
- Features based on the publisher ID of any Google AdSense advertisements embedded in the 77 million pages.
- Features indicative of link exchange based on the UK 2006-05 page-level and host-level web graph.
- Features derived from the URLs in the UK 2006-05 dataset, such as the number of dots, dashes, and digits.
- Features derived from word frequency analysis in the 77 million pages.
- Features based on grouping documents into sets of near-duplicate documents.

We evaluated our best classifier on the 5,622 labelled hosts using ten-fold cross validation. This classifier was constructed using bagging in combination with a C4.5 decision tree. The features used were identified using the Ranker search method and the ChiSquared attribute evaluator.

class	recall	precision
non-spam	96.8%	97.6%
spam	70.5%	80.2%

AIRWeb '07, May 8, 2007 Banff, Alberta, Canada.

2. REFERENCES

[1] C. Castillo, D. Donato, A. Gionis, V. Murdock, F. Silvestri. Know your Neighbors: Web Spam Detection using the Web Topology. *DELIS technical report DELIS-TR-0458*, 2006.

[2] D. Fetterly, M. Manasse, and M. Najork. Spam, Damn Spam, and Statistics: Using statistical analysis to locate spam web pages. *7th International Workshop on the Web and Databases*, 2004.

[3] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting Spam Web Pages Through Content Analysis. *Proceedings of the 15th International World Wide Web Conference (WWW15)*, 2006.

[4] Yahoo! Research: "Web Collection UK-2006". http://research.yahoo.com/ Crawled by the Laboratory of Web Algorithmics, University of Milan, http://law.dsi.unimi.it/. URL retrieved 05 2006.