WEBSPAM-UK2006 Challenge: Hungarian Academy of Sciences Team*

András Benczúr István Bíró Károly Csalogány Miklós Kurucz Tamás Sarlós§
Data Mining and Web Search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{benczur, ibiro, cskaresz, realace, stamas}@ilab.sztaki.hu

ABSTRACT

We use the commercial intent and graph similarity features of our Airweb 2007 and 2006 publications, respectively, in addition to the features of Castillo et al., improving their classification accuracy by 3%. We use stacked graphical learning over the Weka C4.5 classifier.

1. INTRODUCTION

We follow the same methodology as Castillo et al. [5] over the WEBSPAM-UK2006 dataset [4]: we use the *Domain Or Two Humans* classification that introduces additional non-spam domains and gives 10% spam among the 5622 labeled sites. We merge our features with the publicly available ones of [5] and then classify by the C4.5 implementation of the machine learning toolkit Weka.

We use our commercial intent [3] and graph similarity [2] features. We use stacked graphical learning [5] by using weight $1 + \log w$ over the domain graph as well as sites of the same IP address.

2. FEATURES

2.1 Microsoft OCI and Yahoo! Mindset

The Microsoft adCenter Labs Demonstration[1] determines the Online Commercial Intention (OCI) of a URL. Yahoo! Mindset[2] classifies Web pages as either commercial or non-commercial. We assigned a score to each site by issuing an 'inurl:' query to Mindset and then extracted the score corresponding to the site's home page in the returned search engine results. We used raw Mindset values and the logarithm of the OCI probabilities as features for sites where we successfully gathered the values.

2.2 Google AdWords

The AdWords Keyword Tool[3] recommends keywords for a site in the form of a tuple (*group, volume, competition, phrase*). For a query word or phrase, we obtained the estimated average cost per click *CPC* to define the *page cost* of a document by summing up the CPC value of each (known) word occurrence in it and then we average the page costs over each host.

2.3 Google AdSense

Given a site with h pages in the test set, we count the number of pages $p \leq h$ that contain Google AdSense contextual advertisements (`http://www.google.com/adsense`) as well as the total number of Google ads a over the site. Then we assign three features to each host: a, a/p and p/h.

2.4 Spammer search engine success

We define a feature for most popular or competitive queries that describes the extent spammers manage to inject their pages into query top lists over the Hungarian Academy of Sciences Search Engine [1] filled with the WEBSPAM-UK2006 pages. The search engine uses a tf.idf based ranking combined with 25% HostRank scores and increased weights for query words within URL, anchor text, title and additional HTML elements; the engine itself lacks spam filtering.

Given the AdWords scores, we computed the top 1000 hits for each *competition 5 query*. For sites that appeared on the top list we computed and summed up penalties. For position i of a page in the hit list for a query, we obtained the best features by giving score $1/i^2$ for the page. We restricted the location of keyword occurrences to anchors only and rerun the scoring procedure.

We also define the *spam-popularity* weight over queries as follows. For each q of the 10,000 most frequent queries we compute the top 1,000 hits for each query. We give the fraction of spam within labeled[4] (spam / (spam + nonspam)) as weight for q and then compute a weighted penalty sum for each host similarly to the method of competitive queries.

2.5 Additional content features

We computed additional content features including amount of anchor text, fraction of anchor in textual content and length of an anchor text.

2.6 Graph similarity

Of best quality is the fraction of spam within cociting hosts weighted by multiplicity.

3. REFERENCES

[1] A. A. Benczúr, K. Csalogány, E. Friedman, D. Fogaras, T. Sarlós, M. Uher, and E. Windhager. Searching a small national domain—preliminary report. In *Proc. WWW*, 2003.

[2] A. A. Benczúr, K. Csalogány, and T. Sarlós. Link-based similarity search to fight web spam. In *Proc. AIRWeb*, 2006.

[3] A. A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós. Web Spam Detection via Commercial Intent Analysis. In *Proc. AIRWeb*, 2007.

[4] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2), 2006.

[5] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. DELIS Technical report TR-0458, 2006.

*Support from a Yahoo! Faculty Research Grant, project NKFP-2/0024/2005, NKFP-2004 project Language Miner

§T. S. is now with Yahoo! Research, work done while at MTA SZTAKI.

[1] `http://adlab.msn.com/OCI/oci.aspx`

[2] `http://mindset.research.yahoo.com`

[3] `https://adwords.google.com/select/KeywordToolExternal`

[4] For less than 25 labeled hits, we replaced the number of nonspam simply by 25 in order not to overscore due to the large variation.