Web Spam Detection via Commercial Intent Analysis*

András Benczúr István Bíró Károly Csalogány 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, stamas}@ilab.sztaki.hu

ABSTRACT

We propose a number of features for Web spam filtering based on the occurrence of keywords that are either of high advertisement value or highly spammed. Our features include popular words from search engine query logs as well as high cost or volume words according to Google AdWords. We also demonstrate the spam filtering power of the Online Commercial Intention (OCI) value assigned to an URL in a Microsoft adCenter Labs Demonstration and the Yahoo! Mindset classification of Web pages as either commercial or non-commercial as well as metrics based on the occurrence of Google ads on the page. We run our tests on the WEBSPAM-UK2006 dataset recently compiled by Castillo et al. as a standard means of measuring the performance of Web spam detection algorithms. Our features improve the classification accuracy of the publicly available WEBSPAM-UK2006 features by 3%.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Information Storage and Retrieval; I.7.5 [**Document Capture**]: Document analysis

General Terms

Measurement, Experimentation

Keywords

Query popularity, Query monetizability, Commercial intent

1. INTRODUCTION

Identifying and preventing spam is cited as one of the top challenges in web search engines by [16]. As all major search engines incorporate anchor text and link analysis algorithms into their ranking schemes, Web spam appears in sophisticated forms that manipulate content as well as linkage [14]. Spam hunters use a variety of both content [12, 17] and link [15, 9, 20, 3, 2] based features to detect Web spam; a recent measurement of their combination appears in [6].

In this paper we concentrate on identifying pages with a large number of keywords that are either of high advertisement value or highly spammed. As noticed by Gyöngyi and Garcia-Molina [13], most spammers just want financial gain from their activities. In contrast to previous content-based spam features such as distribution, entropy, compressibility targeting the templatic nature of machine generated pages, our features hence try to capture the semantics of spam content. By utilizing external classifiers we also enrich the available training and test data.

We investigate the following features for Web spam detection:

- Online Commercial Intention (OCI) value assigned to an URL in a Microsoft adCenter Labs Demonstration (Section 3.1).
- The Yahoo! Mindset classification of Web pages as either commercial or non-commercial (Section 3.2).
- Google AdWords advertisement keyword suggestions for the sites as well as keyword scores (Section 3.3).
- The distribution of Google AdSense ads over pages of a site (Section 3.4).
- A measure for queries based on spammer success in obtaining high rank for the particular query, measured on our own search engine (Section 3.5).

We run our tests on the WEBSPAM-UK2006 dataset recently compiled by Castillo et al. [5] as a standard means of measuring the performance of Web spam detection algorithms. The baseline decision tree of [6] utilizing content based features achieves an F-measure of 0.610 over our dataset. Inclusion of our new features improves the performance by 3% to an F-measure of 0.641. Similarly, our features boost the F-measure of the content+link based classifier from 0.687 to 0.716 and the stacked graphical learning scheme results from 0.693 to 0.738 as shown in Section 3.6.

1.1 Related results

Ntoulas et al. [17] introduce a number of content based spam features including number of words in page, title, anchor as well as the fraction of page drawn from popular words and the fraction of most popular words that appear in the page. Castillo et al. [6] extend the latter idea by measuring the popularity (frequency) of words in an in-house query log instead of the documents themselves.

*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.

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

Query popularity and monetizability were also recently used to improve cloaking and redirection spam detection. Chellapilla and Chickering [7] aid their cloaking detection method by using the most frequent words from the MSN query log and highest revenue generating words from the MSN advertisement log. As a different method, Wang et al. collect spammer targeted keywords [18] by extracting the most frequent anchor words from spammed forums; they use these keywords for redirection based spam detection.

2. DATASET AND FRAMEWORK

We follow the same methodology as Castillo et al. [6]. We use the WEBSPAM-UK2006 dataset [5] that consists of 71% of the hosts classified as normal, 25% as spam and the remainder 4% as undecided. As in [6] 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 [6] and then classify by the C4.5 implementation of the machine learning toolkit Weka [19].

In addition to the above classification framework of [6] we also evaluate spam filtering by measuring the amount of spam in top hits for queries over the Hungarian Academy of Sciences Search Engine [1]. The search engine uses a tf.idf based ranking combined with 25% HostRank scores [10] and increased weights for query words within URL, anchor text, title and additional HTML elements. The engine itself lacks spam filtering since it is designed primarily for the `.hu` domain that, in our observation, is virtually spam free.

3. ATTRIBUTES AND CLASSIFICATION RESULTS

3.1 Microsoft OCI

Extending Broder's well-known taxonomy of web search [4] the Microsoft adCenter Labs Demonstration available at `http://adlab.msn.com/OCI/oci.aspx` determines the Online Commercial Intention (OCI) of a URL. OCI is described by the probabilities of the URL being commercial-informational, commercial-transactional or non-commercial. The probabilities sum up to 1 and are derived by an SVM based classifier utilizing both the textual content and the HTML tags of the web pages [8]. We have successfully gathered the above mentioned OCI probabilities for the home page of 4995 sites, and failed to do so for the remaining 627 sites, mostly because they were dead when collecting the data in February 2007. Fig. 1 depicts the distribution of the logarithm of commercial-informational scores obtained and shows that spam pages tend to be more commercially oriented.

3.2 Yahoo! Mindset

Yahoo! Mindset (`http://mindset.research.yahoo.com`) classifies Web pages as either commercial or non-commercial. It estimates the commercial nature of a Web page by a value ranging from +2 (most commercial) to −2 (most informational).[1] Pages scored 0 are a balance of commercial and informational. These scores are assigned by a linear SVM based text classifier developed and trained by Y! Research.

We assigned a score to each site in our training and evaluation sample by issuing an 'inurl:' query to Mindset and

[1]Contrary to the Mindset FAQ the actual implementation seems to assign positive values to commercial pages.

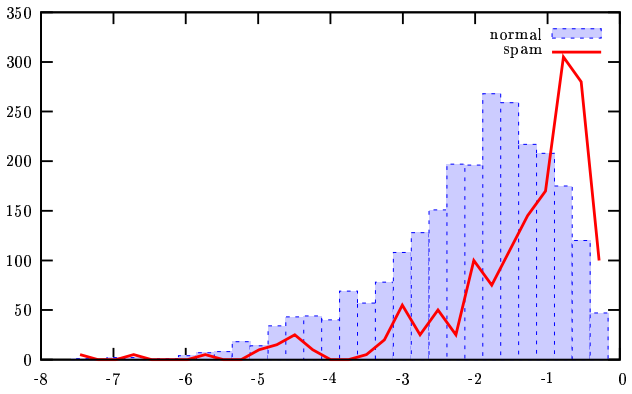

Figure 1: Distribution of the logarithm of OCI commercial-informational score among labeled sites.

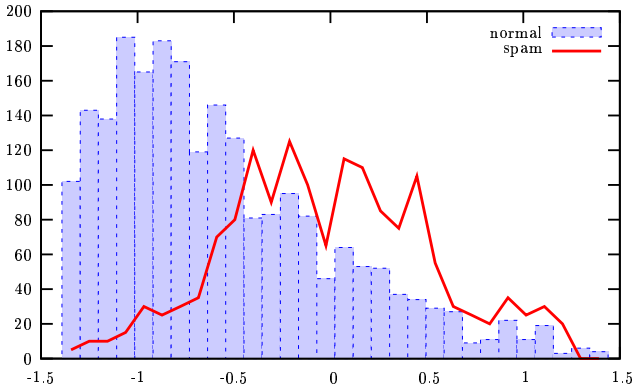

Figure 2: Distribution of Yahoo! Mindset commercial intent score among labeled sites.

then extracted the score corresponding to the site's home page in the returned search engine results.

We have managed to assign a Mindset score to 3170 hosts, the rest were either missing from the current Yahoo! Search index or Mindset failed to classify them. In accordance with Fig. 1, Fig. 2 demonstrates that normal pages are less likely to be commercial in nature as measured by Mindset.

3.3 Google AdWords

AdWords is Google's flagship pay-per-click advertising product (`http://adwords.google.com`). Advertisers bid on keywords and their ads are displayed as sponsored links alongside the organic search results. The AdWords Keyword Tool (`https://adwords.google.com/select/KeywordToolExternal`), that is also available as the API call getKeywordsFromSite(), recommends keywords for a site in the form of a tuple (*group, volume, competition, phrase*). Volume shows the relative amount of users searching for that keyword on Google on a scale 1–5 and advertiser competition shows the relative amount of advertisers bidding on that keyword on the same scale. In addition, for a query word or phrase, we can obtain the following information: estimated average cost per click *CPC*; the *estimated ad positions*, the average position in which the ad may show, expressed in ranges between an upper and lower value. Based on these estimates we 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. The top and middle part of Fig. 3 depicts the distribution of the most discriminative AdWords features.

3.4 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; this latter may be more than h. Then we assign three features to each host: a, a/p (average number of Google ads over pages containing Google ads) and p/h (fraction of pages containing at least one ad).

3.5 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. In contrast to the 10% spam among labeled pages, we see 13% spam among the top 1000 hits of our search engine for popular queries taken from a commercial search engine log. When using highly competitive Google queries, this value increases up to 20%, showing the success of spammers in obtaining high rank in a baseline search engine without spam filtering.

Given the AdWords scores to queries (see Section 3.3) we also obtain features by measuring how well a page fits to the query. Since an excessive study of the possible text based ranking features is beyond the scope of this paper, we simply computed the top 1000 hits for each query using the aforementioned ranking scheme of the Hungarian Academy of Sciences Search Engine [1]. For sites that appeared on the top list we computed penalties that we eventually summed up for all *competition 5 queries*, hence penalizing sites that appear high for several such queries. We have several choices to incorporate the 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. Our feature is finally formed by adding up the $1/i^2$ values of a page for all competition 5 keywords.

Anchor text is perhaps the single most important factor in relevance ranking [11] and hence forms key target for spammers. Amount of anchor text can be used to classify spam [17]. Hence we restricted the location of keyword occurrences to anchors only and rerun the scoring procedure. In Fig. 3, bottom, we graph the distribution of anchors words with advertiser competition value 5 that refer to a given site.

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[2] (spam / (spam + nonspam)) as weight for q and then compute a weighted penalty sum for each host similarly to the method of competitive queries.

3.6 Spam classification accuracy

We evaluate our results by adding the new features to the content based and the content+link based feature sets provided by [5]. We train and test the Weka implementation of the C4.5 decision tree with the same settings as in [6]. We measure accuracy by the F-measure of the spam detection

[2]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.

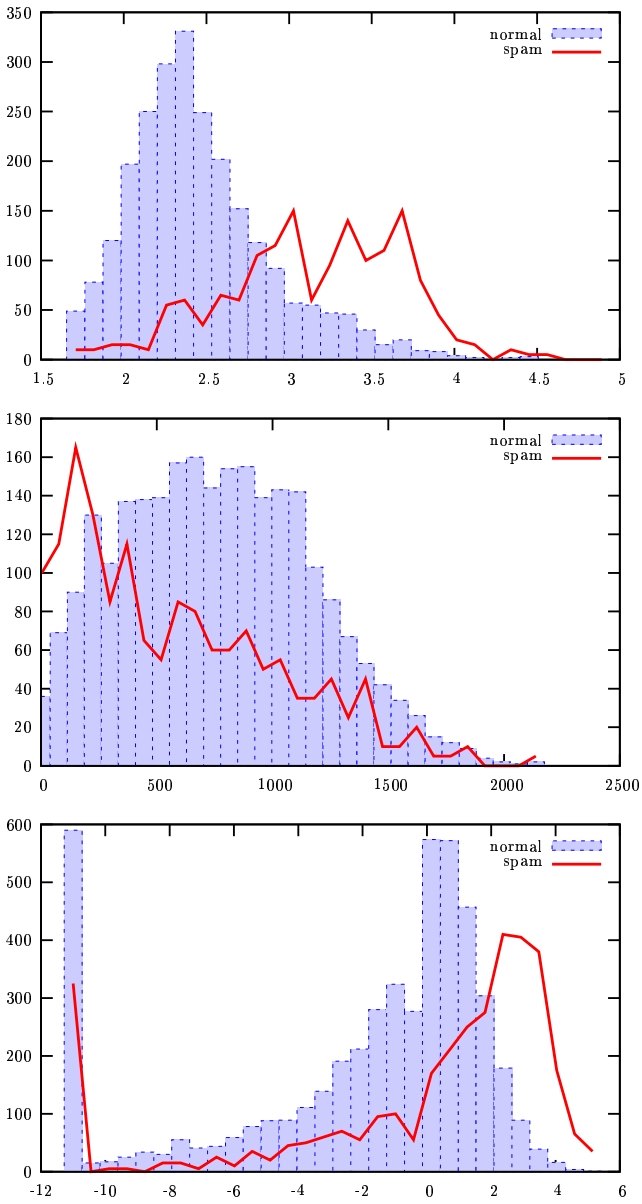

Figure 3: Distribution of Google AdWords based features across labeled spam and nonspam sites. *Top*: average advertiser competition of the site. *Middle*: total estimated upper ad position of the site. *Bottom*: Logarithm of the advertiser competition value 5 term occurrences in anchors to site.

task. Since we could not compute all features for all hosts, we compare our results to the baseline computed on the set of 2292 hosts that have all features we would like to evaluate.

Inclusion of all our new features to the content based features increases the performance of the decision tree by 3% from an F-measure of 0.610 to 0.641. Similarly, our features raise the F-measure of the content+link based classifier from 0.687 to 0.716, which corresponds to 67.1% precision at 76.7% recall. These improvements are statistically significant at a p-value of 19% and 15%, respectively.

Castillo et al. [6] utilize the stacked graphical learning

Feature set	Section	Coverage	C. imp.	C.+L. imp.
OCI	3.1	89%	0.8%	0%
Mindset	3.2	58.6%	0.9%	1.3%
AdWords	3.3	92.5%	1.4%	0.4%
Page cost	3.3	100%	1.1%	0.3%
AdSense	3.4	100%	2.0%	0.5%
Comp. queries	3.5	100%	0.7%	0.4%
C. q. in anchor	3.5	100%	0.4%	0.3%
Spam popular.	3.5	100%	2.2%	1%
All	3	52.4%	3.1%	2.9%

Table 1: Comparison of improvements by feature.

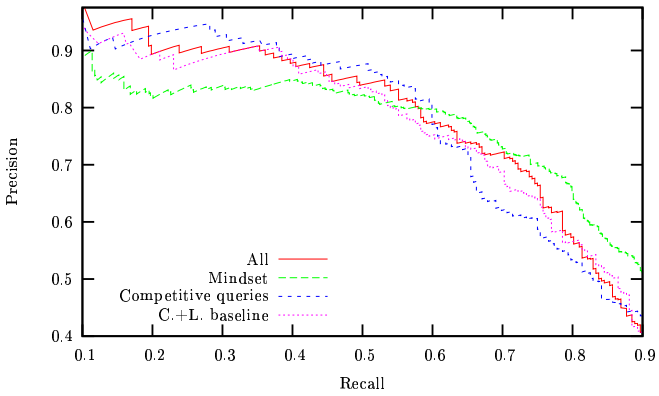

Figure 4: Effect of selected features on accuracy.

method to incorporate the neighborhood of a node in the classification process. According to preliminary experiments, by including our new features we improve the F-measure in this case by over 4% from 0.693 to 0.738.

In Table 1 we analyze the contribution of each feature. We define coverage as the number of sites for which the given feature is available divided 4365, the number of sites for which the content based features of [6] are available. Except for Mindset, all features cover a large fraction of the labeled samples. The fourth and fifth columns list the F-measure improvements achieved by augmenting the content and content+link based features of [6] with a given set of new features. The strongest individual features are Mindset, AdSense, and spam popularity; however neither of them comes close to the combination of all commercial features. The sum the of the OCI commercial estimates is moderately correlated with the Mindset scores, $\rho = 0.54$, which partially explains the weaker individual performance of the former.

Lastly, we depict the precision-recalls curves of the augmented classifiers in Figure 4. Inclusion of most of our features improves precision at lower levels of recall compared to the content+link based classifier of [6]; we plot 'competitive queries' as an example. In contrast, Mindset performs best at high recall and hence the precision curve of the combination of all features stays above the baseline generally.

Conclusion

In this paper we demonstrated the spam filtering power of measuring the commercial intent of a Web page, thus also supporting the observation that most of the Web spammer activities are targeted for financial gains [13].

4. 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, K. Csalogány, T. Sarlós, and M. Uher. SpamRank – Fully automatic link spam detection. In *Proc. AIRWeb*, 2005.

[4] A. Z. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] 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.

[6] 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.

[7] K. Chellapilla and D. M. Chickering. Improving cloaking detection using search query popularity and monetizability. In *Proc. AIRWeb*, pages 17–24, 2006.

[8] H. K. Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li. Detecting online commercial intention (OCI). In *Proc. WWW*, pages 829–837, 2006.

[9] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In *Proc. ECML*, volume 3720 of *LNAI*, pages 233–243, 2005.

[10] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the web frontier. In *Proc. WWW*, pages 309–318, 2004.

[11] R. Fagin, R. Kumar, K. S. McCurley, J. Novak, D. Sivakumar, J. A. Tomlin, and D. P. Williamson. Searching the workplace web. In *Proc. WWW*, 2003.

[12] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics – Using statistical analysis to locate spam web pages. In *Proc. WebDB*, 2004.

[13] Z. Gyöngyi and H. Garcia-Molina. Spam: It's not just for inboxes anymore. *IEEE Computer Magazine*, 38(10):28–34, 2005.

[14] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc. AIRWeb*, 2005.

[15] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. VLDB*, pages 576–587, 2004.

[16] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[17] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. WWW*, pages 83–92, 2006.

[18] Y.-M. Wang, M. Ma, Y. Niu, and H. Chen. Spam double-funnel: Connecting web spammers with advertisers. In *Proc. WWW*, 2007.

[19] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second edition*. Morgan Kaufmann, 2005.

[20] B. Wu, V. Goel, and B. D. Davison. Propagating trust and distrust to demote web spam. In *Workshop on Models of Trust for the Web*, 2006.