Web Spam Filtering in Internet Archives*

Miklós Erdélyi[c,a] András A. Benczúr[a] Julien Masanés[b] Dávid Siklósi[a]
[a]Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{benczur, sdavid}@ilab.sztaki.hu
[b]European Archive Foundation, France
julien@europarchive.org
[c]University of Pannonia
erdelyi@dcs.vein.hu

ABSTRACT

While Web spam is targeted for the high commercial value of top-ranked search-engine results, Web archives observe quality deterioration and resource waste as a side effect. So far Web spam filtering technologies are rarely used by Web archivists but planned in the future as indicated in a survey with responses from more than 20 institutions worldwide. These archives typically operate on a modest level of budget that prohibits the operation of standalone Web spam filtering but collaborative efforts could lead to a high quality solution for them.

In this paper we illustrate spam filtering needs, opportunities and blockers for Internet archives via analyzing several ***crawl snapshots*** and the difficulty of ***migrating filter models*** across different crawls via the example of the 13 `.uk` snapshots performed by UbiCrawler that include WEBSPAM-UK2006 and WEBSPAM-UK2007.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; I.7.5 [**Computing Methodologies**]: Document Capture—*Document analysis*; I.2.7 [**Computing Methodologies**]: Artificial Intelligence—*Natural Language Processing*

General Terms

Web Archival, Information Retrieval

Keywords

Web spam, Document Classification, Time series analysis

1. INTRODUCTION

Current results on Web spam filtering concentrate on the problem of a static crawl and consider the needs of single search companies. The past Web spam challenges[9] as well as most research results [22, 17, 25, 28, 4, 20, 19, 11, 30, 15] to list, in order, the most cited ones[1] were all concentrating on fixed domain crawls with predefined labeled set used for testing and training.

In this paper we consider a very different setup motivated by the needs of Internet preservation. A single archival institution often operates from a low budget that prohibits the development of spam filtering technologies by themselves. Currently 39 archives[2] collaborate under the International Internet Preservation Consortium (IIPC), most of which are national libraries with a primary purpose of national domain preservation crawling. The collaborative and effort sharing nature of the archives is a great advantage compared to the competition among search engines that allows advanced techniques of sharing features and models much beyond the current use of blacklist exchange.

While identifying and preventing spam is a top-priority issue for the search-engine industry [23], so far it is less studied by Web archivists. However, archives are becoming more and more concerned about spam in view of the fact that, under different measurement and estimates, roughly 10% of the Web sites and 20% of the individual HTML pages constitute spam. The above figures directly translate to 10–20% waste of archive resources in storage, processing and bandwidth with a permanent increase that will question the economic sustainability of the preservation effort in the near future.

IIPC is not yet coordinating Web spam filtering efforts, but in a recent survey[3] that we describe in detail in Section 3, 39% of the archives realize spam or fake Web content as a problem in their crawling and capturing process, most of which, in the order of the observed strength of the problem, consist of garbage content, copied content and link farms. In response to another question, they find it difficult to estimate the amount they are able to invest in diminishing spam. Note that results include institutions considering both holistic and selective crawl [24]; selective crawl is less prone to general Web spam but more to spam in social media. Important to emphasize that currently very costly manual filtering is the only option for an archive; for example, a nordic national library is spending 4 man months on filtering after each of its domain crawls.

Spam filtering is essential in Web archives even if we acknowledge the difficulty of defining the boundary between Web spam and honest search engine optimization. Archives may have to tolerate more spam compared to search engines in order not to lose some content misclassified as spam that the users may want to retrieve later. Also they might want to have some representative spam either to preserve an accurate image of the Web or to provide a spam corpus for researchers. In any case, we believe that the quality of an archive with completely no spam filtering policy in use will greatly be deteriorated and significant amount of resources will be wasted

*This work was supported by the EU FP7 project LiWA—Living Web Archives and by grant OTKA NK 72845.

[1]Approximation based on Google Scholar, as of February 2009.

AIRWeb '09, April 21, 2009 Madrid, Spain.

[2]`http://www.netpreserve.org/publications/IIPC_Survey_Report_Public_12152008.pdf`

[3]`http://liwa-project.eu/images/uploads/d1-1.1_requirements_beg_v1.0.pdf`, Annex 1.

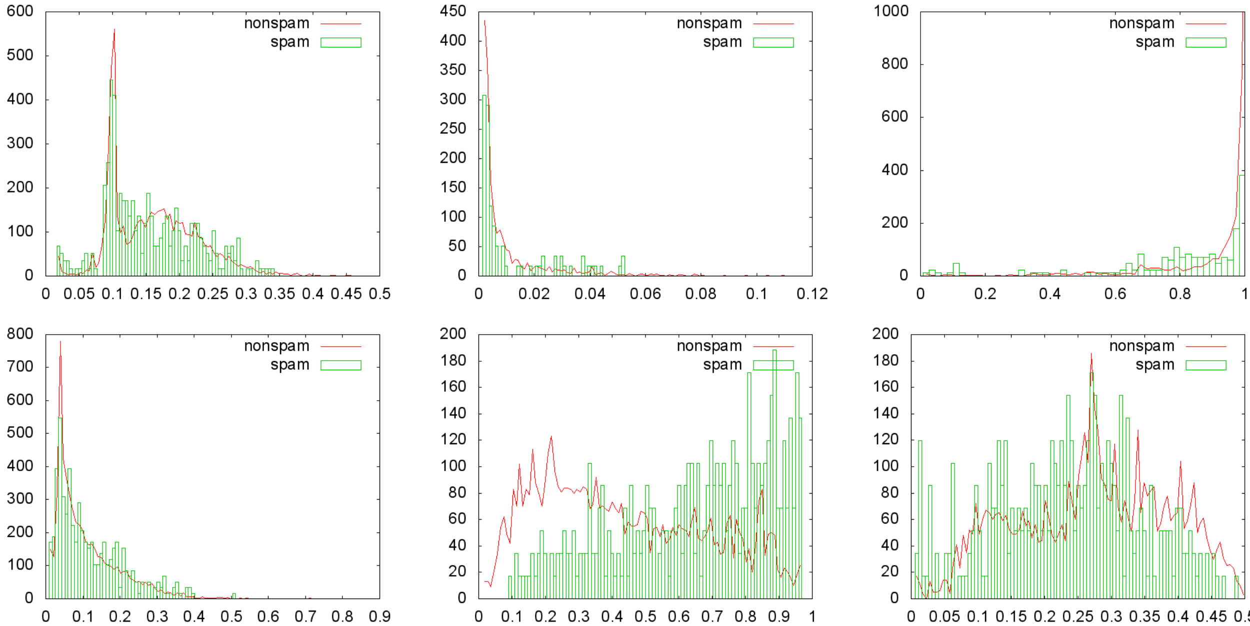

Figure 1: The distribution of feature values across spam and honest pages in the WEBSPAM-UK2007 data set. Top, left to right: cross-snapshot variance of top 500 corpus precision of home page (HST_9) and host-level standard deviation of top 200 corpus recall (STD_84); average host content change in the bag of words model. Bottom, left to right: cross-snapshot variance of fraction of anchor text (HST_4); average and variance of the probability of correctly predicted spamicity in all possible snapshots.

as the effect of Web spam.

The rest of this paper is organized as follows. After listing related results, in Section 2 we describe the time-aware Web spam benchmark collection that we have compiled and the filtering techniques applied so far over the test data. Finally, in Section 3 we review the main results of the survey conducted in Web archives on the estimated effect of spam in their system.

1.1 Related results

As Web spammers manipulate several aspects of content as well as linkage [21], effective spam hunting must combine a variety of content [17, 25, 18] and link [22, 15, 29, 4, 3, 26, 30] based methods. By the lessons learned from the Web Spam Challenges [9], the feature set described in [11] and the bag of words representation of the site content [1] give a very strong baseline with only minor improvements achieved by the Challenge participants. At the current stage of our ongoing work we compute the content features only and use no graph stacking but we plan to use the full power of methods in the future.

Several results investigate the changes of Web content. Earlier results primarily consider this question in conjunction with keeping a search engine index up-to-date [12, 13]. The decay of Web pages and links and its consequences on ranking are discussed in [2, 16]. One main result of Boldi et al. [7] who collected the `.uk` crawl snapshots also used in our results was the efficient handling of time-aware graphs. Closest to our result is the investigation of host overlap, deletion and content dynamics in the same data set by Bordino et al. [8].

2. TIME-AWARE SPAM COLLECTION AND EXPERIMENTS

The main purpose of our experiments is to test the difficulty of the Web archive spam filtering scenarios including time series of snapshots and model transfer for cross-archive collaboration. Our data set consists of the 13 `.uk` snapshots provided by the Laboratory for Web Algorithmics of the Universitá degli studi di Milano together with the Web Spam Challenge labels WEBSPAM-UK2006 and WEBSPAM-UK2007. This invaluable data set of 500GB in WARC 0.19 format consists of the maximum 400 pages per site extract of the original crawls that took 2 weeks to recompile at the original data location and transfer over the network.

The last 12 of the above `.uk` snapshots were analyzed by Bordino et al. [8] who among others observe a relative low URL but high host overlap. The first snapshot (2006-05) that is identical to WEBSPAM-UK2006 was chosen to be left out from their experiment since it was provided by a different crawl strategy. We use this snapshot for testing the possibility of transferring filter models across different crawl strategies.

In order to investigate the usability of the existing labels for the intermediate snapshots we performed overlap measures similar to [8] but considering hosts labeled as spam or honest. The results are available at `http://datamining.sztaki.hu/?q=archive-spam`. We observed fairly high overlap for the last 12 snapshots that justify the usability of models across these crawls with only a moderate expected decay in accuracy. In contrast the first snapshot as well as its labels are apparently of very little use for the later crawls. The first snapshot was fully labeled but due to the different crawl strategy both the fraction of spam is larger than in the WEBSPAM-UK2007 labeling and the decay of these labels is very fast. One possible explanation is that this crawl may have got trapped in certain link farms. We also investigated the overlap of links that is of particular importance for the usability of both the link features and the graph stacking classification.

2.1 Temporal spam features

We define new features based on the time series of the "public" content and link features [10]. The distribution of sample features

AUC	07	08	09	10	11	12	01	02	03	04	05
06	784	783	762	763	734	747	744	720	760	731	712
07		727	771	724	757	777	746	720	762	777	717
08			862	839	885	864	851	886	869	852	785
09				791	788	814	783	830	794	747	763
10					813	786	757	797	783	787	781
11						850	788	826	830	776	763
12							753	724	736	750	793
01								782	778	854	837
02									800	828	765
03										822	782
04											768

Table 1: The AUC of model transfer across the 2006-06...2007-05 snapshots, multiplied by 1000 for better readability. The earlier snapshot is used for training and the later is for testing. Classification is by C4.5 over public content features [10].

for spam and honest hosts are shown in Fig. 1.

First we define ***centralized*** versions of each feature to make one snapshot comparable to another as follows. For very skew distributed features such as degree we switch to using the logarithm. Then from each feature we subtract the average over the entire snapshot and use the value as the new centralized feature.

Next we compute the ***variance*** of all features across the snapshots. We use a 5-month training and testing period that starts in the 2006-08 snapshot the earliest in order to avoid possible noise due to the possible initial stabilization of the crawl parameters. Variance is simply computed over the centralized values of the same feature over all snapshots in question. As a key observation, we realize that if a feature has large variance for a host, then this particular feature and host pair is less reliable for classification.

Due to the variance of its features, certain hosts turn out to be less reliable for classification. We define ***stability*** as the variance of the probability of making a correct prediction when classifying a given host as part of a heldout set defined by 5-fold partitioning of the training set.

We also analyze the fraction of content change over the site. We compute the bag of words for the union of all pages in the host and compute the Jaccard and cosine similarity across the crawl snapshots. Finally we aggregate by average, maximum and variance to form new features for each host.

In our classifier ensemble we split features into related sets and for each we use the best fitting classifier. These classifiers are then combined by random forest, a method that, in our crossvalidation experiment, outperformed logistic regression suggested by [14]. We used the classifier implementations of the machine learning toolkit Weka [27].

2.2 Spam filtering results

For the purposes of our experiments we have computed all the public Web Spam Challenge content features of [10]. The link feature generation is under progress and hence we are considering classification based on the content features only. All classification below are by C4.5 over these content features.

In our first experiment we consider model transfer across different crawl snapshots. When using WEBSPAM-UK2006 with very different crawl strategy, the model performs poor despite of the fact that the training set here consists of all 10,662 labeled hosts of WEBSPAM-UK2007, as seen in the last column of Table 2. For the remaining snapshot pairs we observe little impact of the time difference. For the results in Table 1 we define the training and test sets as the collection of hosts that appear in all of the 12 last crawl snapshots. We also repeat the experiment for classifying newly appeared hosts in the 2007-05 snapshot.

Next we define new features for WEBSPAM-UK2007 based on

Setup	Challenge	New host	2006 → 2007
Training set size	1,201	4,000	10,662
Public content [10]	0.753	0.699	0.730
BOW	0.619	-	-
Stability	**0.776**	-	-
Variance	0.618	-	-

Combinations	Challenge
Content + BOW	0.729
Content + stability	0.766
Content + variance	0.726
Content + BOW + stability + variance	**0.777**

Table 2: AUC results for the WEBSPAM-UK2007 data set and combination of classifiers. BOW denotes featues based on content change in the bag of words model of the host. Training and test sets are defined as follows: Challenge denotes the Web Spam Challenge 2008 testing labels, new host denotes those newly appeared in 2007-05, and finally for 2006 → 2007 training was on WEBSPAM-UK2006 and testing on WEBSPAM-UK2007.

Blog comment spam	20% (2)
Link farms	50% (5)
Copied content	60% (6)
Garbage content	70% (7)

Table 3: Distribution of 10 responses to question "If you do meet spam during capturing, of what type is that spam?".

the earlier snapshots. The results for combining the content features with their variance and the classifier stability are summarized in Table 2.

3. WEB ARCHIVES SURVEY RESULTS

In a survey conducted as part of the LiWA—Living Web Archives project user requirement analysis we received invaluable response from more than 20 archival institutions related to their opinion, existing and planned policies related to Web spam. For the question "Is spam or fake Web content a problem in your crawling and capturing process?" 9 out of 23 responses (39%) were positive and only one respondent considered no problem caused by spam even in the future. The types of spam they have already met is summarized in Table 3.

More important is the planned actions to prevent archives from spam. While archives often consider spam as a necessary part of the present state of the Web content that may even need to be preserved, several institutions apply blacklists and filters as summarized in Table 4.

In contrast to the observed problem, the resources are low on spam filtering. For the question "If you undertake actions to diminish the spam problem in the Web archive of your institute, can you estimate how much you invest in this?" we had only 8 responses

We drop pages with spam or fake content.	18,20% (2)
We drop sites with spam or fake content.	45,50% (5)
We apply filters to avoid such noise.	54,50% (6)
After capturing we manually correct the crawl.	27,30% (3)
We see no options to avoid noise.	27,30% (3)

Table 4: Distribution of 11 responses to question "If spam has impact on your Web archiving process, what actions do you undertake?".

with three considering it "difficult to estimate". Other responses were "I would spend perhaps 3 or 4 days creating lists of seeds to filter out of the forthcoming crawl."; "10 minutes – 1 hour per site."; "We use 2-5 minutes per website when going through the list of potential spam sites.". It is important to consider the problem addressed in one response: "We do not do anything to edit captured content. We foresee that this would not scale, and that it would invite questions about the archive's authenticity. This is something that content owners we interviewed were very concerned about - that their captured content be protected from alteration."

Conclusion

With illustration over the 100,000 page WEBSPAM-UK2007 snapshot of the `.uk` domain, we have reported ongoing work for preventing Web spam in Internet archives, a key element for the economic sustainability of the preservation effort. The implementation of filtering has a promising start by taking advantage of time depth that archives provide and the non-competitive environment that allows collaboration. By our findings we may conclude that the classification of newly appeared hosts, the use of time series features and the transformation of filter models for a different crawl open new research questions and may serve as tasks for a future Web Spam Challenge [5].

Acknowledgment

To Sebastiano Vigna, Paolo Boldi and Massimo Santini for providing us with the UbiCrawler crawls [6, 7]. In addition to them, also to Illaria Borodino, Carlos Castillo and Debora Donato for discussions on the WEBSPAM-UK data sets [8] and ideas on a possible new Web Spam Challenge based on periodic recrawls.

4. REFERENCES

[1] J. Abernethy, O. Chapelle, and C. Castillo. WITCH: A New Approach to Web Spam Detection. In *Proc. of the 4th Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[2] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins. Sic transit gloria telae: Towards an understanding of the Web's decay. In *Proceedings of the 13th World Wide Web Conference (WWW)*, pages 328–337. ACM Press, 2004.

[3] A. A. Benczúr, K. Csalogány, and T. Sarlós. Link-based similarity search to fight Web spam. In *Proc. of the 2nd Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2006.

[4] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. SpamRank – Fully automatic link spam detection. In *Proc. of the 1st Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[5] A. A. Benczúr, M. Erdélyi, J. Masanés, and D. Siklósi. Web spam challenge proposal for filtering in archives. In *AIRWeb '09: Proc. of the 5th Int. Workshop on Adversarial Information Retrieval on the Web*. ACM Press, 2009.

[6] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler: A scalable fully distributed Web crawler. *Software: Practice & Experience*, 34(8):721–726, 2004.

[7] P. Boldi, M. Santini, and S. Vigna. A Large Time Aware Web Graph. *SIGIR Forum*, 42, 2008.

[8] I. Bordino, P. Boldi, D. Donato, M. Santini, and S. Vigna. Temporal evolution of the uk Web. In *Workshop on Analysis of Dynamic Networks (ICDM-ADN'08)*, 2008.

[9] C. Castillo, K. Chellapilla, and L. Denoyer. Web spam challenge 2008. In *Proc. of the 4th Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[10] 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):11–24, December 2006.

[11] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the Web topology. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 423–430, 2007.

[12] J. Cho and H. Garcia-Molina. The evolution of the Web and implications for an incremental crawler. In *The VLDB Journal*, pages 200–209, 2000.

[13] J. Cho and H. Garcia-Molina. Synchronizing a database to improve freshness. In *Proceedings of the International Conference on Management of Data*, pages 117–128, 2000.

[14] G. Cormack. Content-based Web Spam Detection. In *Proc. of the 3rd Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2007.

[15] I. Drost and T. Scheffer. Thwarting the nigritude ultramarine: Learning to identify link spam. In *Proc. of the 16th European Conference on Machine Learning (ECML)*, volume 3720 of *Lecture Notes in Artificial Intelligence*, pages 233–243, 2005.

[16] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the Web frontier. In *Proc. 13th International World Wide Web Conference (WWW)*, pages 309–318, New York, NY, USA, 2004. ACM Press.

[17] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics – Using statistical analysis to locate spam Web pages. In *Proceedings of the 7th International Workshop on the Web and Databases (WebDB)*, pages 1–6, Paris, France, 2004.

[18] D. Fetterly, M. Manasse, and M. Najork. Detecting phrase-level duplication on the world wide Web. In *Proceedings of the 28th ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, Salvador, Brazil, 2005.

[19] Z. Gyöngyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen. Link spam detection based on mass estimation. In *Proc. 32nd Int. Conference on Very Large Data Bases (VLDB)*, 2006.

[20] Z. Gyöngyi and H. Garcia-Molina. Link spam alliances. In *Proc. 31st Int. Conference on Very Large Data Bases (VLDB)*, 2005.

[21] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc. 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2005.

[22] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web spam with TrustRank. In *Proc. 30th International Conference on Very Large Data Bases (VLDB)*, pages 576–587, 2004.

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

[24] J. Masanès. *Web archiving*. Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006.

[25] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam Web pages through content analysis. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, pages 83–92, Edinburgh, Scotland, 2006.

[26] PR10.info. BadRank as the opposite of PageRank, 2004. `http://en.pr10.info/pagerank0-badrank/` (visited June 27th, 2005).

[27] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2nd edition, June 2005.

[28] B. Wu and B. D. Davison. Identifying link farm pages. In *Proceedings of the 14th International World Wide Web Conference (WWW)*, pages 820–829, Chiba, Japan, 2005.

[29] 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*, Edinburgh, Scotland, 2006.

[30] B. Wu, V. Goel, and B. D. Davison. Topical TrustRank: Using topicality to combat Web spam. In *Proceedings of the 15th International World Wide Web Conference (WWW)*, 2006.