Web Spam Challenge Proposal for Filtering in Archives*

András A. Benczúr[a] Miklós Erdélyi[c,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

In this paper we propose new tasks for a possible future Web Spam Challenge motivated by the needs of the archival community. The Web archival community consists of several relatively small institutions that operate independently and possibly over different top level domains (TLDs). Each of them may have a large set of historic crawls. Efficient filtering would hence require (1) enhanced use of the time series of domain snapshots and (2) collaboration by transferring models across different TLDs. Corresponding Challenge tasks could hence include the distribution of crawl snapshot data for feature generation as well as classification of unlabeled new crawls of the same or even different TLDs.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; I.7.5 [**Computing Methodologies**]: Document Capture—*Document analysis*

General Terms

Web Archival, Information Retrieval

Keywords

Web spam, Document Classification, Challenge, Evaluation

1. INTRODUCTION

Web spam filtering know-how became widespread with the success of the Adversarial Information Retrieval Workshops since 2005 that host the Web Spam Challenges since 2007. In order to initiate collaboration between the Web archival and the Spam filtering communities, we intend to provide time-aware Web spam benchmark data sets for future Web Spam Challenges.

Web Spam Challenges were organized with the purpose of identifying and comparing Machine Learning methods for automatically labeling Web hosts represented as graphs with feature vectors over nodes. These past challenges as well as most research results in the area (see references within [8]) concentrate on the problem of a single crawl with training and testing labels.

In this paper we propose a very different setup for a future Web Spam Challenge motivated by the needs of Internet archives. We describe new training and testing scenarios. New features may be generated by considering the temporal change of several crawl snapshots of the same domain [10, 7, 8]. In addition by the needs of collaboration across different archival institutions we may also provide training labels over one TLD and request prediction over a fully or partly unlabeled different domain.

By the distributed operation of the archival community and the high importance of the collaborative efforts we face several problems. When sharing knowledge across different domains, not just the primarily used languages will differ, but also the linkage and the crawl strategies in use. The difficulty may be balanced by the fact that each archive contains several crawl snapshots of the same domain. These snapshots can be used for generating features based on changes and time series. For example by analyzing content change, parked domains can be more efficiently caught. By the nature of the collaborative efforts and the access to several snapshots we define the following research questions.

1. Classification by using features based on time series, content change, appearance and disappearance of pages and sites.
2. Classification of newly appeared hosts.
3. Using a spam classification model compiled over an earlier crawl to filter the current crawl.
4. Using a spam classification model compiled over a completely different crawl of different strategy and possibly even over a different top level domain.

2. TASKS AND DATA SETS

In this section we overview existing data sets as well as the possible additional crawls and part of the required manual assessment labels provided by the European Archive Foundation that will be available for the purposes of the Challenge. We elaborate on the possible tasks of the existing data and initiate discussion on planning the crawls in accordance with the needs of the proposed tasks.

Currently 13 monthly `.uk` UbiCrawler crawls [3, 4] are available for testing purposes as a courtesy of Sebastiano Vigna, Paolo Boldi and Massimo Santini. The first 2006-05 and the last 2007-05 snapshots consist of WEBSPAM-UK2006 and WEBSPAM-UK2007 along with labels distributed as in Table 1. The first snapshot (2006-05) is provided by a different crawl strategy and by our preliminary tests [8] this data set indicates the difficulty of transferring filter models across different crawl strategies.

In order to take the temporal change of the corpus into account, we may compile a new crawl of the `.uk` domain around the la-

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

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

	WEBSPAM-UK2006	WEBSPAM-UK2007
normal	8,123	5,709
spam	2,113	344
undecided	426	376
total labeled	10,662	6,429
total hosts	10,662	114,529

Table 1: The number of hosts in Web Spam Challenge data.

beled sites of the WEBSPAM-UK2007 data set. Note that a careful selection procedure is required since the Internet Archive crawl of the `.uk` domain currently consists of over 2M sites, an amount that exceeds the capacity of the possible Challenge participants and organizers. In the future two tasks are possible:

1. New site classification. Assessors label sites that are not present in the 12 original `.uk` snapshots; only the existing WEBSPAM-UK2007 labels are available for training.
2. Temporal feature generation. From a sequence of periodic recrawls generate time-aware spam features as well as perceive changes in the behavior of certain sites.

The data sets required for the first task are relatively easy to compile. For the second task more crawls are needed together with a specific scope that limits the volume of data to be processed by participants.

Label sets for different snapshots of hosts already labeled in WEBSPAM-UK2007 may also help in learning possible ownership changes such as transforms from an honest site into a parked domain abused by spammers. For this purpose new labels for hosts with a large fraction of their content and linkage changed could be assessed. In a possible scenario one host may appear with real content and gather some in-links. The host then becomes *parked* because the owner gives up operation. Pages from such a host remain in search engine caches presenting valuable entry points for spammer farms. When major search engines realize the change and blacklist the host, spammers may give up their operation over this domain or the domain may even reopen with honest content again.

The European Archive Foundation currently starts crawls of the `.eu` domain prone to severe spammer activities. The Archive has resources to assess a sample of hosts in this domain that can be used for evaluation purposes in two scenarios.

- A single snapshot is made available with testing labels only. The task is to use a model compiled over WEBSPAM-UK2007 for this `.eu` data set.
- Multiple snapshots are made available along with both training and testing labels with the main task consisting of handling multilingualism over the domain. In this scenario another TLD crawl would be necessary for the previous task.

3. EXISTING AND EXPECTED FILTERING TECHNOLOGIES

Various top-level or otherwise selected domains have different level of spammer activities; Ntoulas et al. [11] give a comparison that show major differences among national domains and languages of the page. In general, however, a very similar spammer behavior is observed in all major TLD so that we may accept findings of the Web Spam Challenge participants conclusive for most domains.

The current state of the art in spam filtering is summarized in the best performing systems [6, 1, 9] of the Web Spam Challenges. Most results either use the tf.idf vectors or the so-called "public" feature sets of [5]. The Web Spam Challenge 2008 best result [9] used ensemble undersampling while for earlier challenges, best performances were achieved by a semi-supervised version of SVM [1] and text compression [6]. In a recent result [2] we find very strong performance of the SVM classifier alone over simply the tf.idf vectors, a fact that could question the importance of spam feature generation. However, tf.idf will clearly fail when we move to a TLD with different dominant language (e.g. `.de`) or mixed language (`.eu`).

In summary, the new challenge tasks may take stronger use of language independent features such as the "public" ones [5] instead of the more traditional text classification techniques (tf.idf, SVM etc.). New features based on time series [10, 7, 8] as well as normalization methods across different snapshots and TLDs are the expected outcome of the proposed tasks.

4. OPEN QUESTIONS

The distribution of data sets for a possible new Spam Challenge appears challenging itself. The current compilation of the 12 `.uk` snapshots has a maximum number of 400 pages from each of the approximately 100,000 hosts. One such snapshot has a size near .5TBytes which implies the data sets could only be distributed on disks by mail. According to the current estimate and seed list of the Internet Archive, the `.uk` domain is expected to consist of over 2M hosts while the `.eu` of 3.2M. A selection procedure is needed to include only part of the hosts, in particular if multiple snapshots are to be distributed. As an alternate solution, only feature sets such as "public" [5] can be made available; in this case a precompiled set of content change features based e.g. on [7] should also be compiled.

5. REFERENCES

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

[2] I. Bíró, D. Siklósi, J. Szabó, and A. A. Benczúr. Linked latent dirichlet allocation in web spam filtering. In *AIRWeb '09: Proc. 5th int. workshop on Adversarial information retrieval on the web*, 2009.

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

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

[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):11–24, December 2006.

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

[7] N. Dai, B. D. Davison, and X. Qi. Looking into the past to better classify web spam. In *AIRWeb '09: Proc. 5th int. workshop on Adversarial information retrieval on the web*, 2009.

[8] M. Erdélyi, A. A. Benczúr, J. Masanés, and D. Siklósi. Web spam filtering in internet archives. In *AIRWeb '09: Proc. 5th int. workshop on Adversarial information retrieval on the web*, 2009.

[9] G. Geng, X. Jin, and C. Wang. CASIA at WSC2008. In *Proc. 4th Int. Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

[10] Y. joo Chung, M. Toyoda, and M. Kitsuregawa. A study of web spam evolution using a time series of web snapshots. In *AIRWeb '09: Proc. 5th int. workshop on Adversarial information retrieval on the web*, 2009.

[11] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly. Detecting spam web pages through content analysis. In *Proc. of the 15th Int. World Wide Web Conference (WWW)*, pages 83–92, 2006.