Web Spam Filtering in Internet Archives

Miklós Erdélyi, András A. Benczúr, Julien Masanés, Dávid Siklósí

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

Categories and Subject Descriptors

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

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 as well as most research results to list, in order, the most cited 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 collaborate under the International Internet Preservation Consortium (IPC), 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, 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.

IPC is not yet coordinating Web spam filtering efforts, but in a recent survey that we describe in detail in Section 3, 39% of the archives realize spam or fake 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; 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.

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 Technologies of the Università 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
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 WEBSHAM-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 WEBSHAM-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 WEBSHAM-UK2007 based on 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

<table>
<thead>
<tr>
<th>Setup</th>
<th>Challenge</th>
<th>New host</th>
<th>2006 → 2007</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set size</td>
<td>1.201</td>
<td>4.000</td>
<td>10.662</td>
</tr>
<tr>
<td>Public content</td>
<td>0.753</td>
<td>0.699</td>
<td>0.730</td>
</tr>
<tr>
<td>BOW</td>
<td>0.619</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stability</td>
<td>0.776</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Variance</td>
<td>0.618</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: AUC results for the WEBSHAM-UK2007 data set and combination of classifiers. BOW denotes features 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 WEBSHAM-UK2006 and testing on WEBSHAM-UK2007.

<table>
<thead>
<tr>
<th>Combinations</th>
<th>Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content + BOW</td>
<td>0.729</td>
</tr>
<tr>
<td>Content + stability</td>
<td>0.766</td>
</tr>
<tr>
<td>Content + variance</td>
<td>0.726</td>
</tr>
<tr>
<td>Content + BOW + stability + variance</td>
<td>0.777</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Type of Spam</th>
<th>Group</th>
<th>Count</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blog comment spam</td>
<td>20%</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Link farms</td>
<td>50%</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Copied content</td>
<td>60%</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Garbage content</td>
<td>70%</td>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

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