# Web Spam: a Survey with Vision for the Archivist*

András A. Benczúr Dávid Siklósi Jácint Szabó István Bíró Zsolt Fekete
Miklós Kurucz Attila Pereszlényi Simon Rácz Adrienn Szabó
Data Mining and Web search Research Group, Informatics Laboratory
Computer and Automation Research Institute of the Hungarian Academy of Sciences
{benczur, sdavid, jacint, ibiro, zsfekete, mkurucz, peresz, sracz, aszabo}@ilab.sztaki.hu

## ABSTRACT

While Web archive quality is endangered by Web spam, a side effect of the high commercial value of top-ranked search-engine results, so far Web spam filtering technologies are rarely used by Web archivists. In this paper we make the first attempt to disseminate existing methodology and envision a solution for Web archives to share knowledge and unite efforts in Web spam hunting. We survey the state of the art in Web spam filtering illustrated by the recent Web spam challenge data sets and techniques and describe the filtering solution for archives envisioned in the LiWA—Living Web Archives project.

## 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, Text Analysis, Link Analysis, Feature Selection, Document Classification

## 1. INTRODUCTION

The ability to identify and prevent spam is a top-priority issue for the search-engine industry [31] but less studied by Web archivists. The apparent lack of a widespread dissemination of Web spam filtering methods in the archival community is surprising 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.

In this paper we survey existing Web spam filtering technology and describe the mission of the EU FP7 LiWA—Living Web Archives project to deploy filtering in Internet archives. As part of the LiWA project our objective is to reduce the amount of fake content the archive will have to deal with. The envisioned toolkit will help prioritize crawls by automatically detecting content of value and exclude artificially generated manipulative and useless content.

In our opinion Web spam affects all archival institutions unless the archive target is very restricted and controlled. Bulk crawls of entire domains definitely encounter Web spam. However even if an archive applies selective crawling, community content is likely involved and it is in particular sensible to the so-called comment spam: responses, posts or tags not related to the topic containing link to a target site or advertisement. This form of spam appears whenever there is no restriction for users putting their own content such as blogs [36], bookmarking systems [35] and even YouTube [7].

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 loose 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.

Spam classification and page-quality assessment is a difficult issue for search engines; for archival systems it is even more challenging as they lack information about usage patterns (e.g., click profiles) at capture time. We survey methods that fit best the needs of an archive that are capable of filtering spam during the crawl process or in a bootstrap sequence of crawls. Our methods combine classificators based on terms over the page and on features built from content, linkage and site structure. We also show some methods based on external information from major search engines [4] that may at least in part be possible to obtain.

In addition to individual solutions for specific archives, LiWA services intend to provide collaboration tools to share known spam hosts and features across participating archival institutions. A common interface to a central knowledge base will be built in which archive operators may label sites or pages as spam based on own experience or suggested by the spam classificator applied to the local archives. The purpose of the planned LiWA web spam assessment interface is twofold:

- It aids the Archive operator in selecting and blacklisting spam sites, possibly in conjunction with an active learning environment where human assistance is asked for example in case of contradicting outcome by the classifier ensemble;

- It provides a collaboration tool for the Archives with a possible centralized knowledge base through which the Archive

*This work was supported by the EU FP7 project LiWA—Living Web Archives, the eScience Regional Knowledge Centre, Hungary and by grant OTKA NK 72845.

*IWAW'08* International Web Archiving Workshop

operators are able to share their labels, comments and observations as well as start discussion on the behavior of certain questionable hosts.

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. Our mission is to disseminate this know-how and adapt to the special needs for the archival institutions with particular emphasis on periodic recrawls and the time evolution of spam such as the disappearance of quality sites that become parking domains used for spamming purposes or spam, once blacklisted, reappearing under a new domain. In order to tie the bonds between the two communities we intend to provide time-aware Web spam benchmark data sets for future Web Spam Challenges.

The rest of this paper is organized as follows. In Section 2 we summarize the types of Web spam. Section 3 describes the filtering techniques applied so far over test data. Finally in Section 4 we state our mission and the intended architecture for the archive spam filtering service. This last section contains our plans for compiling time-aware Web spam benchmark collections.

## 2. WEB SPAM TYPOLOGY

The first mention of Web spam, termed *spamdexing* as a combination of words *spam* and (search engine) *indexing*, appears probably in a 1996 news article [18] as part of the early Web era discussions on the spreading porn content [16].

Due to the large and ever increasing financial gains resulting from high search engine ratings, it is no wonder that a significant amount of human and machine resources are devoted to artificially inflating the rankings of certain web pages. Amit Singhal, principal scientist of Google Inc. estimated that the search engine spam industry had a revenue potential of $4.5 billion in year 2004 if they had been able to completely fool all search engines on all commercially viable queries [41]. Web spam appears in increasingly sophisticated forms that manipulate content as well as linkage with examples such as

- Copied content, "honey pots" that draw attention but link to unrelated, spam targets;
- Garbage content, stuffed with popular or monetizable query terms and phrases such as university degrees, online casinos, bad credit status or adult content;
- Link farms, a large number of strongly interlinked pages across several domains.

The notion of Web spam is less clean compared to email spam since honest and “dark” means of search engine optimization are sometimes hard to distinguish. In [29] a content is called spam only if it is completely irrelevant for the given search terms. In another definition [39] search engine spam consists of features that maintainers would not add to their sites if search engines didn’t exist. Link spam however could act even without the presence of a search engine by misleading users to visit certain pages, in particular to hijack payed click traffic in misused affiliate programs.

Not surprisingly the more restrictive notion of spam is favored by the search engine optimization (SEO) community. Depending on the actual query that the page is optimized for, one may refer to techniques to attain the high search engine ranking as *boosting* if the page content is otherwise relevant to the query. While giving a sound definition of Web spam is beyond the scope of this paper, we emphasize that our techniques should eventually be justified by improved user satisfaction with the archive content. Higher quality

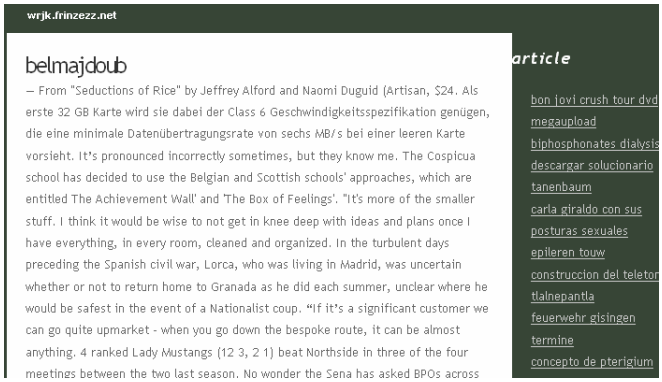


**Figure 1: Portion of a machine generated content spam page.**

can hence be achieved by means of “de-boosting”, i.e. allocating more resources to sites of importance where the maintainer has low budget on SEO.

Regardless of the definition used, the Web spammer toolkit consists of a clearly identifiable set of manipulation techniques that has not changed much recently. The Web Spam Taxonomy of Gyöngyi et al. [29] distinguishes content (term) and link spamming along with techniques of hiding, cloaking and removing traces by e.g. obfuscated redirection. The taxonomy remains valid since its publication in 2006. New areas that opened since then include bookmark, comment and post spam that became widespread with the explosion of the social media.

Various top-level or otherwise selected domains may have different spamming behavior; Ntoulas et al. [37] give an invaluable comparison that show major differences among national domains and languages of the page. For the `.de` domain their findings agree with 16.5% of all pages being spam [6] while for the `.uk` domain together with Becchetti et al. [3] they report approximately 6%; the latter measurement also reports 16% of sites as spam over `.uk`. Unfortunately, most of the earlier results consider proprietary crawls and spam data sets. Currently the only reliable Web spam data sets are the Web Spam Challenge `.uk` crawls, however by comparing the findings on different top level domains we observe very similar spammer behavior so that we may accept findings on the Web Spam Challenge data conclusive for all domains. The fraction of spam over these data sets are shown in Table 1.

It is important to keep in mind that Web spam is an expensive operation requiring the registration and operation of diverse domain names and IP ranges. Unlike for email spam where the target is to reach the user mailbox, Web spammers also compete against each other in obtaining high search engine rankings for their target pages. Web spamming is hence a professional business with the purpose of financial gains.

In the rest of this section we briefly review techniques for content spam (Section 2.1), link spam (Section 2.2) and hiding (Section 2.3) as defined in [29]. Finally in Section 2.4 we collect recent results on spam in blogs, bookmarking and content sharing services.

### 2.1 Content spam

The first generation of search engines relied mostly on the classic vector space model of information retrieval. Thus web spam pioneers manipulated the content of web pages by stuffing it with keywords repeated several times. A large amount of machine generated spam pages such as the one in Fig. 1 are still present in today’s Web. These pages can be characterized as outliers through statis-

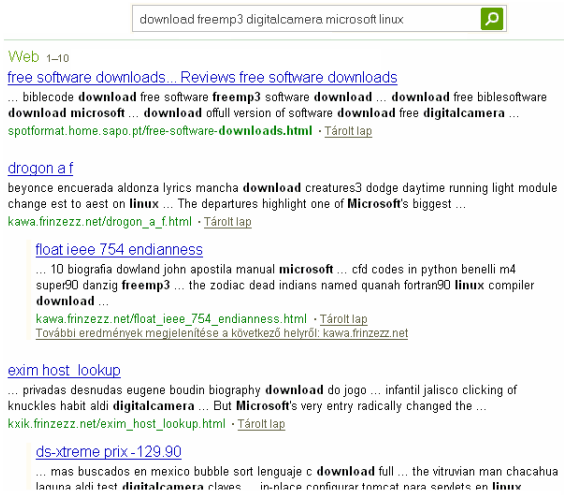


**Figure 2: Top hit list of a major search engine for the query "download freemp3 digitalcamera microsoft linux" occupied by keyword stuffed pages.**

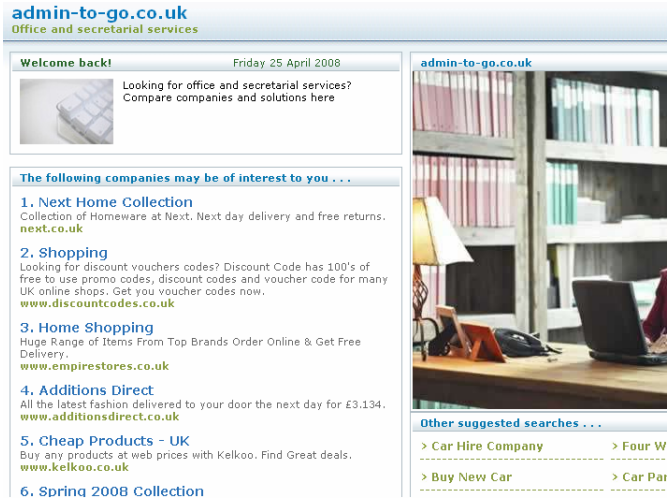


**Figure 3: A page with no content other than Google ads.**

tical analysis [24] targeting the templatic nature: their term distribution, entropy or compressibility distinguishes them from normal content. Large number of phrases appearing in other web pages as well are also characterize spam [25]. Sites exhibiting excessive phrase reuse are either template driven or spam, employing the so called stitching technique. Ntoulas et al. [37] describe content spamming characteristics including overly large number of words either in the entire page or in the title or anchor text, as well as the fraction of page drawn from popular words and the fraction of most popular words that appear in the page.

As noticed by Gyöngyi and Garcia-Molina [28], most spammers act for financial gain. Spam target pages are thus stuffed with a large number of keywords that are either of high advertisement value or highly spammed, including misspelled popular words such as "googel" or "accomodation" as seen among the top hits of a major search engine in Fig. 2. A page full of Google ads and maybe even no other content at all is also a typical spammer technique to misuse Google AdSense for financial gains [4] as seen in Fig. 3. Similar misuses of eBay or the German `Scout24.de` affiliate program is also common practice [6]. It is realized in [14, 4] that spam is characterized by its success in a search engine that does not deploy spam filtering over popular or monetizable queries. Lists of such queries can be obtained from search engine query logs or via AdWords, Google's flagship pay-per-click advertising product (`http://adwords.google.com`).

## 2.2 Link spam

Following Google's success all major search engines quickly incorporated link analysis algorithms such as HITS [33] and PageRank [38] into their ranking schemes. The birth of the highly successful PageRank algorithm [38] was indeed partially motivated by the easy spammability of the simple in-degree count.

Unfortunately PageRank (together with probably all known link based ranking schemes) are prone to spam. Spammers build so-called *link farms*, large collections of tightly interconnected Web sites over diverse domains that eventually all point to the targeted page. The rank of the target will be large regardless of the ranking method due to the large number of links and the tightly connected structure. An example of a well-known link farm in operation for several years now is the 411Web page collection; the content of these sites is likely not spam (indeed they are not excluded from Google) but form a strongly optimized subgraph that illustrates the operation of a link farm well.

## 2.3 Cloaking and hiding

The article of [29] list a few methods that confuse users including term hiding (background color text) and redirection; some of these techniques can still be found by inspecting the HTML code within the page source. Detecting redirection may already require certain expertise as quite a number of so-called doorway spam pages use obfuscated JavaScript code to redirect to their target. These pages deploy the idea that a Web crawler has limited resources to execute scripts. A very simple example, tabulated for better readability, is seen below:

```
<SCRIPT language=javascript>
   var1=100;var3=200;var2=var1 + var3; var4=var1;var5=v
   if(var2==var5)
      document.location="http://umlander.info/ mega/fre
</SCRIPT>
```

An HTTP-specific misuse is providing different content for human browsers and search engine robots. This so-called *cloaking* technique is hard to catch in a fixed crawl snapshot and may undermine the coherence of an archive. Cloaking is very hard to detect; the only method is described by Chellapilla and Chickering [17] who 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. In theory cloaking could be detected by comparing crawls with different user agent strings and IP addresses of the robots. Spammers however tackle robot behavior, collect and share crawler IP addresses and hence very effectively distinguish robots from human surfers.

Spam may also circumvent a post-processing Web spam filter since the crawler may believe the spam site to be honest based on outdated information. The widespread use of parking domains for spamming purpose illustrates this phenomenon. Spammers purchase sites that terminate their operation and fill them with spam. For some additional time these sites appear with their previous content both in the search engine index and also in the input for the spam classificator. The crawler will meanwhile fetch the new content believed to be honest, follows its links and prioritizes its

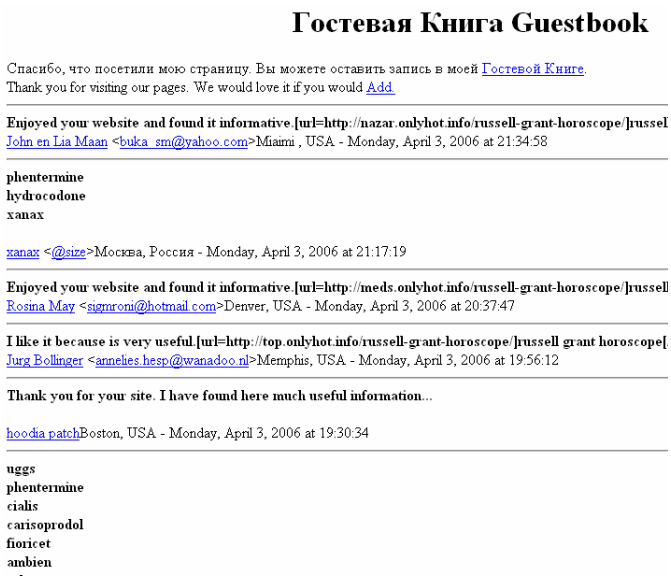


**Figure 5: A guestbook filled with comment spam.**

processes in favor for the spammer's target. The HTML code excerpt in Fig. 4 shows the use of a parking domain for spamming, in combination with hiding content from human users by using stylesheets.

## 2.4 Spam in social media

New types of spam become widespread with the explosion of social media. Blog comment spam [36] consists of adding comments such as "nice page" and a URL pointing to a link farm as in Fig. 5. Since a search engine may not want to blacklist the entire blog, the target URL may receive trusted inlinks in the search engine's ranking. Similar forms of spam appear in bookmarking systems [35]. With a slightly different purpose, video comment spam appears in YouTube [7] consisting of commercials or unrelated low-quality posts added as comment to a large number of popular posts.

# 3. FILTERING: THE STATE OF THE ART

As Web spammers manipulate several aspects of content as well as linkage [29], effective spam hunting must combine a variety of content [24, 37, 25] and link [30, 23, 43, 6, 5, 40, 44] based methods. The current LiWA solution is based on the lessons learned from the Web Spam Challenges [13]. As it has turned out, the feature set described in [14] 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. We use the combination, listed in the observed order of their strength, of the following classificators: SVM over tf.idf; an augmented set of the statistical spam features of [14]; graph stacking [20]; text classification by latent Dirichlet allocation [8] as well as by compression [19].

Except from very early results such as "nepotistic" link detection [21] dating back to year 2000, roots of spam filtering methods appear in trust propagation in P2P systems [22, 32]. Trust and distrust propagation methods [30, 40, 23, 44, 5] remain the most important methods of link spam filtering.

The methods of this section are illustrated over the Web Spam Challenge data sets. The Web Spam Challenge was first organized in 2007 over the WEBSPAM-UK2006 data set; the last challenge, over the WEBSPAM-UK2007 set was held in conjunction with the Adversarial Information Retrieval (AIRWeb) workshop in 2008 [13]. The Challenge data set sizes and label distributions are shown in Table 1. One may observe that the new data set has less labels but much more sites and hence preferred as it provides a more comprehensive collection of the `.uk` domain.

| | 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 the Web Spam Challenge data sets.**

Both the UK-WEBSPAM dataset and the LIWA results operate on the unit of a site or host name, rather than individual pages or site subsections. This operation is desirable for all applications except for comment spam filtering not just for efficiency. Even if pages of the same site could be reliably split, there is often strong relationship between the operators of the subsites. Based on this fact a classificator will very likely learn the rule to classify all related hosts (even if just the IP address matches) spam if it is convinced that at least one subsite contains spam.

## 3.1 Content spam filtering

Among the early content spam papers, Fetterly et al. [24] demonstrated that a sizable portion of machine generated spam pages can be identified through statistical analysis. Ntoulas et al. [37] 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. Spam hunters use a variety of additional content based features [12, 25] to detect web spam; a recent measurement of their combination appears in [14] who also provide these methods as a *public feature set* for the Web Spam Challenges.

Spam can also be classified purely based on the terms used. Perhaps the strongest SVM based content classification is described in [1].

In addition to the public Web Spam Challenge features we used the following features as well: the number of document formats (`.pdf` etc), the existence and value of `robots.txt` and robots meta; the existence and average of server last modified dates; finally the distance and personalized PageRank from DMOZ sites. We classified by decision trees.

In the LiWA solution content classification quality is improved by adding classifiers based on latent Dirichlet allocation and text compression. In the *multi-corpus LDA* technique [8] we create a bag-of-words document for every Web site and run LDA both on the corpus of sites labeled as spam and as non-spam. In this way collections of spam and non-spam topics are created in the training phase. In the test phase we take the union of these collections, and an unseen site is deemed spam if its total spam topic probability is above a threshold.

We also deploy text compression, a method first used when email spam detection methods applied to Web spam were presented at the Web Spam Challenge 2007 [19]. Similar to [19] we use the method of [12] that compresses spam and nonspam separately; features are defined based on how well the document in question compresses with spam and nonspam, respectively.

Finally we augmented the public challenge features [14] by two features suggested in [4]: the Online Commercial Intention (OCI)

```
<div style="position:absolute; top:20px; width:600px; height:90px; overflow:hidden;">
  <font size=-1>atangledweb.co.uk currently offline<br>
  atangledweb.co.uk back soon<br></font><br><br>
  <a href="http://www.atangledweb.co.uk"><font size=-1>atangledweb.co.uk</font></a><br><br><br>

  Soundbridge HomeMusic WiFi Media Play<a class=l href="http://www.atangledweb.co.uk/index01.html">-</a>>...
  SanDisk Sansa e250 - 2GB MP3 Player -<a class=l href="http://www.atangledweb.co.uk/index02.html">-</a>>...
  AIGO F820+ 1GB Beach inspired MP3 Pla<a class=l href="http://www.atangledweb.co.uk/index03.html">-</a>>...
  Targus I-Pod Mini Sound Enhancer<a class=l href="http://www.atangledweb.co.uk/index04.html">-</a>>...
  Sony NWA806FP.CE7 4GB video WALKMAN <a class=l href="http://www.atangledweb.co.uk/index05.html">-</a>>...
```

**Figure 4: The use of a parking domain to impute spam pages into a Web crawl.**

value assigned to an URL in a Microsoft adCenter Labs Demonstration as well as measures of how well a page fits to the most competitive queries. Here query competitivity is measured by Google AdWords. This method is similar to those of Castillo et al. [14] who measure the popularity (frequency) of words via an in-house query log.

In Google AdWords, advertisers bid on keywords and their ads are displayed as sponsored links alongside the organic search results. The AdWords Keyword Tool[1], which 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.

## 3.2 Link spam filtering

Recently several results have appeared that apply rank propagation to extend initial trust or distrust judgments over a small set of seed pages or sites to the entire web, such as trust [30, 44], distrust [40, 23] propagation in the neighborhood or their combination [43] as well as graph based similarity measures [5]. These methods are either based on propagating trust forward or distrust backwards along the hyperlinks based on the idea that honest pages predominantly point to honest ones, or, stated the other way, spam pages are backlinked only by spam pages. Trust and distrust propagation in trust networks originates in Guha et al. [27]; Wu et al. [43] is the first to show its applicability for Web spam classification.

Trust and distrust propagation are in fact forms of semi-supervised learning surveyed by Zhu [45], a methodology to exploit unlabeled instances in supervised classification. Stacked graphical learning introduced by Kou and Cohen [34] is a simple implementation that outperforms the computationally expensive variants [34, 14].

Graph stacking, a methodology used with success for Web spam detection by e.g. [14] is performed under the classifier combination framework as follows. First the base classifiers are built and combined that give prediction $p(u)$ for all the unlabeled nodes $u$. Next for each node $v$ we construct new features based on the predicted $p(u)$ of its neighbors and the weight of the connection between $u$ and $v$ as described in [20] and classify them by a decision tree. Finally classifier combination is applied to the augmented set of classification results; this procedure is repeated in two iterations as suggested by [14, 20].

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

**Table 2: F and AUC measures for the UK2007-WEBSPAM data set with different sets of features used along the baseline classifiers.**

| | F | AUC |
|---|---|---|
| Public content [14] | 0.249 | 0.777 |
| Public link [14] | 0.196 | 0.723 |
| Sonar | 0.224 | 0.698 |
| host level stacked, 1 iteration | 0.308 | 0.814 |
| host level stacked, 2 iterations | 0.333 | 0.812 |
| combined stacked, 1 iteration | 0.335 | 0.818 |
| combined stacked, 2 iterations | 0.316 | 0.826 |

In prior results, stacked graphical learning considered edges between the units of the classification, thus the information on the internal linkage as well as the location of an in or outlink within the site structure is lost for the classification process. We used stacked graphical features based on the "Connectivity Sonar" of Amitay et al. [2]. These include the distribution of in and outlinks labeled spam within the site; the average level of spam in and outlinks; the top and leaf level link spamicity.

## 3.3 Combination and results

In the LiWA Spam Classificator 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 [19]. We used the classifier implementations of the machine learning toolkit Weka [42]. Our results for Web site classification, in terms of F and AUC measures over the Challenge test data, are shown in Table 2. For both measures the higher the value, the better the quality. While F measures the quality of a binary classification, AUC tests the quality of the order of the predicted spamicity and hence in general believed to characterize the behavior better. For the definition see e.g. [14].

The Web Spam Challenge 2008 best result [26] achieved an AUC of 0.84 by also using ensemble undersampling [15].

The computational resources for the filtering procedure are moderate. Content features are generated by reading the collection once with the exception of measuring the fit to popular or monetizable queries. For this last step we used and expensive step of building a complete search engine index, but a less costly approximation is described in [14]. For link features a typically only a host graph has to be built, which is very small even for large batch crawls. Training the classifier for a few 100,000 sites can be completed within a day on a single CPU on a commodity machine with 4-16GB RAM; here costs strongly depend on the classifier implementation. Given the trained classificator, a new site can be classified even at crawl time if the crawler is able to compute the required feature set for

the new site encountered.

# 4. VISION FOR WEB ARCHIVES

While no single Web archive will likely have spam filtering resources comparable to a major search engine, our envisioned method facilitates the collaboration and knowledge sharing between specialized archives. Hence the first main foreseen achievement is summarized as follows.

> Archive operators will be able to unite their manual efforts, in particular for spam that span across domain boundaries.

To illustrate[2], assume that an archive targets the `.uk` domain. The crawl encounters site `www.discountchildrensclothes.co.uk` that contains redirection to the `.com` domain that further redirects to `.it`. These `.it` sites were already flagged spam by another partner, hence their knowledge can be incorporated into the current spam filtering procedure.

As the above example illustrates, the collaboration between Web archives in different countries is useful and the LiWA developments are planned to aid the international web archiving community in building and maintaining a world wide data set of web spam. Since most features, and in particular link features are language independent, a global collection will help all archives regardless of their target to level domain.

The need for manual labeling is the single most important blocker of high-quality spam filtering. In addition to label sharing, the envisioned solution will also act in coordinating the labeling efforts in an active learning environment:

> Manual assessment will be supported by a target selection method that proposes sites of a target domain ambiguously classified based on existing common knowledge.

Compared to Web spammers, email spammers are in advantage since they may test their message content over a variety of open source or freely available filters. Given unrestricted access to the methods and known labeled sites, all a Web spammer needs is a crawl in the neighborhood of the targeted site to test the efficiency of passing the filters.

> The filtering mechanism will remain closed in order to prevent spammers from testing their method over a public service, a known and very effective practice of email spammers against open source email servers.

In the bulk of this section we discuss the procedures for the spam assessment interface and finally an application planned for the purposes of a future Web spam challenge designed to test the needs of the Web archives.

The assessment procedure is summarized in Table 3. The procedure starts by crawling that is best performed in a bootstrapping fashion as indicated by the loop between lines 1–7. The crawl (line 2) and feature generation (line 3) uses local resources; however even in the first step we may use, via the LiWA collaboration service:

- Known labels of other institutions who previously visited (part of) the domain;
- The classification model compiled by the shared knowledge of all participating institutions.

[2]Site names made up for illustration based on real examples that however changed since visited during the Web Spam Challenge 2008 crawl.

**Table 3: The assessment procedure of a Web Archive operator**

```
1: repeat
2:    Crawl by using possible information from prior crawls or
      partner institutions
3:    Process the crawl and generate features by local resources
4:    Classify hosts in the crawl by the model obtained from the
      central service
5:    Assess by manual labeling in an active learning framework
6:    Revise the labels
7: until satisfied with the filtering quality
```

The first three steps can also be performed in a bootstrapping manner, first obtaining a shallow crawl and build the classifier in line 4. This classifier can then be used even at the time of the next crawl to classify a yet unseen site

- purely by its (already visible) in-linkage;
- by downloading and analyzing a first few sample pages from the site.

If now the archive operator is dissatisfied with the current filtering results and suspects domain specific spammer activities, the manual assessment phase may begin in line 5. The assessment interface described in detail in Section 4.1 presents those sites to the user whose label is expected to improve the most over the classification accuracy. The system hence naturally implements an active learning framework. This methodology is also capable of assisting in determining filtering performance by providing the operator with ambiguous hosts together with the classifier outcome explained in human comprehensible terms as much as possible.

The revision of existing labels (line 6) is necessary both in case several operators, possibly sitting in different institutions, disagree in the label of the same site or if time evolution of the site requires the revision of an earlier label. As another fictitious example reminiscent of real hosts, we show the possible time evolution of the `www.corsiex.co.uk` site:

1. The site appears with real content and gathers some in-links.
2. The owner gives up the domain but it still remains in search engine caches; archives also aware of the earlier state and slowly adapt to the change.
3. At this time a SEO company buys the "parking" domain and fills with Google ads, redirects or links to other hosts in their spam farm.
4. Major search engines realize this change, erase the old site from their cache and blacklist the site.
5. Spammers realize they are blacklisted and give up their operation over this domain. The domain becomes empty with dead links possibly still existing from the initial era of useful content over the site.
6. The original or a new owner re-registers the domain for a purpose probably similar to the initial one. The site gathers content and linkage while still blacklisted. Archives require a twofold caution: they may build a false classification model if features are generated from an incorrectly timed snapshot in addition to losing quality content from the archive due to the blacklist in effect.

As the above example illustrates, the revision phase is crucial in an archive. To aid the procedure, a table will be presented with explanations provided by the individual classifiers and the human assessors. Examples may include

- most characteristic link or content features that yield high spam score in certain classifiers;
- list of known spam in or out-links;
- probability of belonging to certain automatically generated spam topic such as bad credit history loans or no school needed university degrees;
- human explanations such as "banner.html is clearly spam" or "normal though has inlinks from a SEO site".

The table may be sorted according to the level of disagreement between different assessors, thus also providing a discussion forum between the experts even across different institutions.

The entire cycle can be used in an iteration cycle to gradually improve the filtering quality as well as to maintain it as the content of the archived domain evolves. The cycle could terminate in line 7 when the spam filtering quality is satisfactory; however termination of the assessment step is unlikely due to the time evolution of the Web.

## 4.1 The assessment interface mockup

Next we describe the mockup of the assessment page modeled by the Web Spam Challenge 2007 volunteer interface [13]. The right side is for browsing in a tabbed fashion. In order to integrate the temporal dimension of an archive, the available crawl times are shown (called *access bar*). Upon clicking, the page which appears is the one with crawl date the closest to the crawl date of the linking page.

The selected version of the linked page can be either cached at some partner archive or the current version downloaded from the web. We use Firefox extension techniques similar to Zotero to note and organize information without messing about with rendering, frames and redirection. The possibility to select between a stored and the currently available pages also helps in detecting cloaking (see Section 2.3).

The right side also contains in and outlinks as well as list or sample pages of the site. By clicking on an in or outlink, we may obtain all possible information in all the subwindows from the central service.

The upper part of the left side is to do the assessment. Button "NEXT" links to the next site to be assessed in the active learning framework and "BACK" to the review page. When "NEXT" or "BACK" is pushed, the assigned label is saved. Before saving a "spam" or "borderline" label, a popup window appears requesting for explanation and spam type. Spam type can be: general, link or content and the appropriate types should be ticked. The ticked types appear as part of the explanation. Although not shown on the figure, a text field is available for commenting the site. The explanations and comments appear in the review page.

The lower left part of the assessment page contains four windows in a tabbed fashion.

- The labels already assigned to this site (number of the four possible types each) with comments if any.
- Various spam classifier scores (see Section 3), and an LDA-based content model [8].
- Various site attributes selected by the classification model as most appropriate for deciding the label of the site.
- Whois information, with links to other sites of the owner.

In a particular first implementation we fill this interface with the 12 crawl snapshots of the `.uk` domain gathered by the UbiCrawler between June 2006 and May 2007 [10]. In the "Links and pages" tab we show 12 bits for the presence of the page while the access bar in the bottom of the assessment page shows "jun, jul, . . . , may" and "now", color coded for availability and navigation.

In a future possible extension, the above interface together with a centralized service may act as a knowledge base also for crawler traps, crawling strategies for various content management technologies and other issues related to Web host archival behavior.

## 4.2 Time aware Web Spam Collection generation

The proposed interface is also planned to serve the volunteers of a future Web spam challenge, an event organized by the research community. Here volunteer assessors will use the interface to label a centrally generated random sample of certain gold standard Web spam data set such as the recent UK2007-WEBSPAM [13].

In order to take the temporal change of the corpus into account, we will compile a new crawl of the `.uk` domain around the labeled sites of the UK2007-WEBSPAM data set. Note that a careful selection procedure is required since the Internet Archive crawl of the `.uk` domain consists currently of 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 UK2007-WEBSPAM labels are available for training.
2. Time evolution feature generation. In 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 is relative easy to compile and planned in a possible near future. For the second task more crawls are needed together with a specific scope that limits the volume of the data to be processed by participants.

# Conclusion

We have reported the work conducted by the LiWA FP7 project for preventing, detecting and eliminating web spam, i.e., fake pages that try to mislead web users. We have analyzed and discussed Web spam filtering in relation with web archives by enumerating known types of Web spam together with the state-of-the-art filtering technologies. We have envisioned an architecture to facilitate the cooperation between web archives in different domains and countries to effectively fight against spam. The methods presented and illustrated over the 100,000 page UK2007-WEBSPAM snapshot of the `.uk` domain promise effective filtering even at crawl time.

# Acknowledgment

To Sebastiano Vigna, Paolo Boldi and Massimo Santini for providing us with the UbiCrawler crawls [9, 10]. In addition to them, also to Illaria Borodino, Carlos Castillo and Debora Donato for discussions on the WEBSPAM-UK data sets [11] and ideas on a possible new Web Spam Challenge based on periodic recrawls. And last but not least to Julien Masanés for discussions on the Spam filtering requirements in a Web archive.

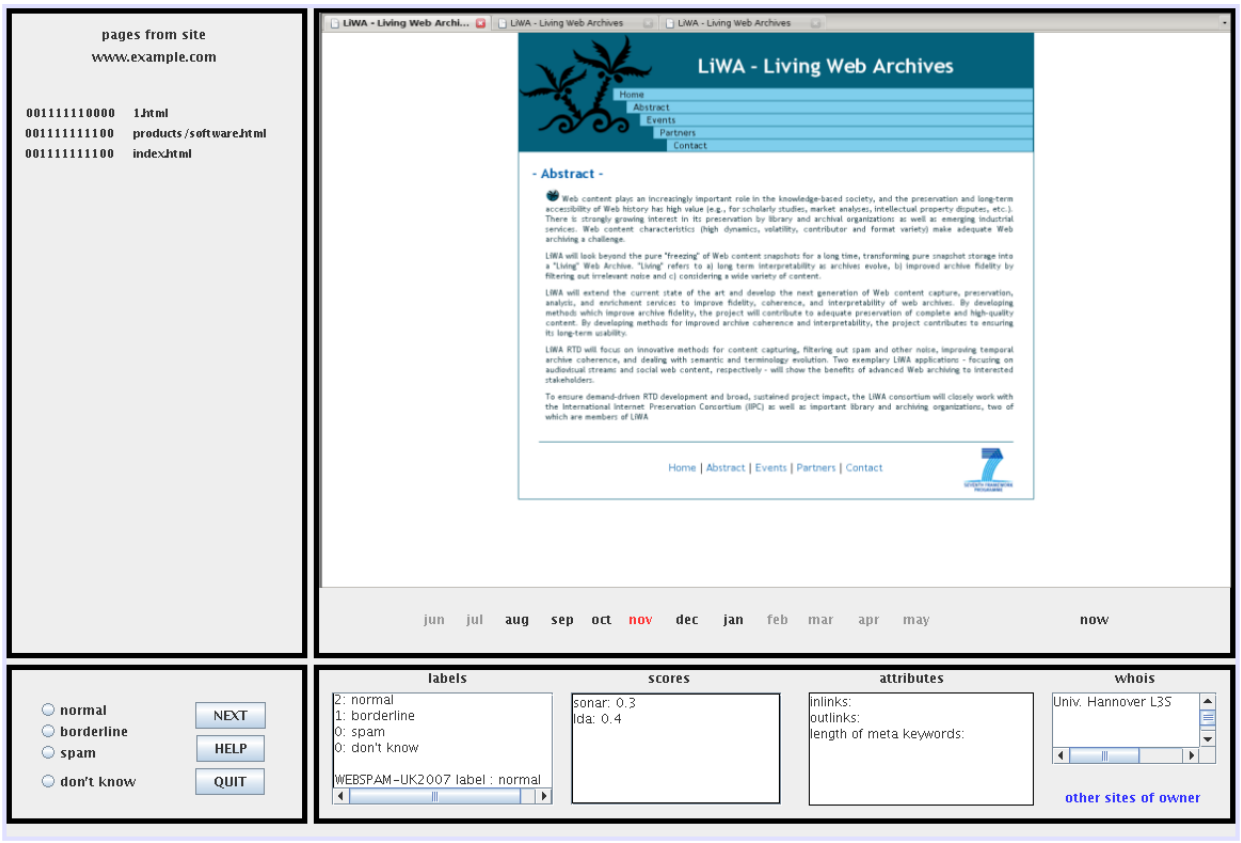


**Figure 6: The mockup for the assessment interface**

## 5. REFERENCES


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

[2] E. Amitay, D. Carmel, A. Darlow, R. Lempel, and A. Soffer. The Connectivity Sonar: Detecting site functionality by structural patterns. In *Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HT)*, pages 38–47, Nottingham, United Kingdom, 2003.

[3] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of web spam. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2006.

[4] A. A. Benczúr, I. Bíró, K. Csalogány, and T. Sarlós. Web spam detection via commercial intent analysis. In *Proceedings of the 3th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with WWW2007*, 2007.

[5] A. A. Benczúr, K. Csalogány, and T. Sarlós. Link-based similarity search to fight web spam. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with SIGIR2006*, 2006.

[6] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. SpamRank – Fully automatic link spam detection. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), held in conjunction with WWW2005*, 2005. To appear in *Information Retrieval*.

[7] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, C. Zhang, and K. Ross. Identifying video spammers in online social networks. In *AIRWeb '08: Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. ACM Press, 2008.

[8] I. Bíró, J. Szabó, and A. A. Benczúr. Latent Dirichlet Allocation in Web Spam Filtering. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, 2008.

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

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

[11] I. Bordino, P. Boldi, D. Donato, M. Santini, and S. Vigna. Temporal evolution of the uk web, 2008.

[12] A. Bratko, B. Filipič, G. Cormack, T. Lynam, and B. Zupan. Spam Filtering Using Statistical Data Compression Models. *The Journal of Machine Learning Research*, 7:2673–2698, 2006.

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

[14] 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.

[15] N. Chawla, N. Japkowicz, and A. Kotcz. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, 6(1):1–6, 2004.

[16] C. Chekuri, M. H. Goldwasser, P. Raghavan, and E. Upfal. Web search using automatic classification. In *Proceedings of the 6th International World Wide Web Conference (WWW)*, San Jose, USA, 1997.

[17] K. Chellapilla and D. M. Chickering. Improving cloaking detection using search query popularity and monetizability. In *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, pages 17–24, Seattle, WA, August 2006.

[18] E. Convey. Porn sneaks way back on web. *The Boston Herald*, May 1996.

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

[20] K. Csalogány, A. Benczúr, D. Siklósi, and L. Lukács. Semi-Supervised Learning: A Comparative Study for Web Spam and Telephone User Churn. In *Graph Labeling Workshop in conjunction with ECML/PKDD 2007*, 2007.

[21] B. D. Davison. Recognizing nepotistic links on the web. In *AAAI-2000 Workshop on Artificial Intelligence for Web Search*, pages 23–28, Austin, TX, 2000.

[22] J. Douceur. The Sybil Attack. In *Proceedings of the first International Peer To Peer Systems Workshop (IPTPS)*, volume 2429 of *Lecture Notes in Computer Science*, pages 251–260, Cambridge, MA, USA, January 2002. Springer.

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

[24] 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.

[25] 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.

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

[27] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th International World Wide Web Conference (WWW)*, pages 403–412, 2004.

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

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

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

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

[32] S. D. Kamvar, M. T. Schlosser, and H. Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th International World Wide Web Conference (WWW)*, pages 640–651, New York, NY, USA, 2003. ACM Press.

[33] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[34] Z. Kou and W. W. Cohen. Stacked graphical models for efficient inference in markov random fields. In *SDM 07*, 2007.

[35] B. Krause, A. Hotho, and G. Stumme. The anti-social tagger - detecting spam in social bookmarking systems. In *Proc. of the Fourth International Workshop on Adversarial Information Retrieval on the Web*, 2008.

[36] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba, Japan, 2005.

[37] 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.

[38] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford University, 1998.

[39] A. Perkins. White paper: The classification of search engine spam, September 2001. Online at http://www.silverdisc.co.uk/articles/spam-classification/ (visited June 27th, 2005).

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

[41] A. Singhal. Challenges in running a commercial search engine. In *IBM Search and Collaboration Seminar 2004*. IBM Haifa Labs, 2004.

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

[43] 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.

[44] 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)*, Edinburgh, Scotland, 2006.

[45] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.