Identifying Web Spam with User Behavior Analysis[1]

Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru
State Key Lab of Intelligent technology & systems,
Tsinghua National Laboratory for Information Science and Technology,
CS&T Department, Tsinghua University, Beijing, 100084, China P.R.
yiqunliu@tsinghua.edu.cn

ABSTRACT
Combating Web spam has become one of the top challenges for Web search engines. State-of-the-art spam detection techniques are usually designed for specific known types of Web spam and are incapable and inefficient for newly-appeared spam. With user behavior analyses into Web access logs, we propose a spam page detection algorithm based on Bayesian Learning. The main contributions of our work are: (1) User visiting patterns of spam pages are studied and three user behavior features are proposed to separate Web spam from ordinary ones. (2) A novel spam detection framework is proposed that can detect unknown spam types and newly-appeared spam with the help of user behavior analysis. Preliminary experiments on large scale Web access log data (containing over 2.74 billion user clicks) show the effectiveness of the proposed features and detection framework.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Search process, H.3.4 [**Systems and Software**]: Performance evaluation

General Terms
Experimentation

Keywords
Spam detection, Web search engine, User behavior analysis

1. INTRODUCTION
With the explosive growth of information on the Web, search engines become more and more important in people's daily lives. In China, there are more than 120 million Internet users, among which 74.8% use search engines frequently and 84.5% regard using search engines as a major way to find newly-appeared information [1]. Although a search engine usually returns thousands of results for a certain query, most search engine users only view the first few pages in result lists according to [2]. As a consequence, the ranking position has become a major concern of internet service providers.

In order to get "an unjustifiably favorable relevance or importance score for some Web page, considering the page's true value"[3], various kinds of Web spam techniques were designed to mislead search engines. In 2006, it is estimated that about one seventh of English Web pages are spam and these spam lead to great obstacle in users' information acquisition process [10]. Therefore, spam detection is regarded as a major challenge for Web search service providers [4].

State-of-the-art anti-spam techniques usually make use of Web page features, either content-based or hyper-link structure based, to construct Web spam classifiers. In this spam detection framework, when a certain kind of Web spam appears in search engine results, anti-spam engineers examine the characteristics of this kind of spam and design specific strategies to identify it. However, once one kind of spam is detected and banned, the spammers will develop new Web spam techniques instantly. Since the beginning of search engines' wide adoption in the late 1990s, Web spam has evolved from term spamming, link spamming to current hiding and JavaScript spamming techniques. Although machine learning based methods have shown their superiority for being easily adapted to newly-developed spam, these approaches still require researchers to provide specific spam page's features and build up suitable training sets.

This kind of anti-spam framework has caused many problems in the development of Web search engines. Anti-spam has become an ever-lasting process but it can only detect Web spam types which have caused severe loss and have drawn anti-spam engineers' attention. It is quite difficult for anti-spam techniques to be designed and implemented in time because when the engineers are aware of a certain spam type, it has succeed in attracting much users' attention.

Compared with the prevailing approaches, we propose a different anti-spam framework: the User Behavior-oriented Web Spam Detection framework. Web spam attempts to deceive search engine ranking algorithm instead of meeting Web user's information needs as ordinary pages. Therefore, the user-visiting patterns of Web spam pages differ from ordinary Web pages. By collecting and analyzing large-scale user-access data of Web pages, we find several user behavior features of spam pages. These features are used to develop an anti-spam algorithm to identify Web spam in a timely, effective, and type-independent manner.

In summary, the contributions of the paper are:

1. We propose a Web spam detection framework in which spam sites are identified because of their deceitful motivation instead of their content/hyper-link appearance.
2. We introduce three features developed from user behavior pattern analyses and these features can identify spam Web sites from ordinary sites timely and effectively.
3. We design a learning-based approach to combine the proposed user-behavior features and compute the likelihood that the Web sites are spam.
4. We construct a user access corpus of over 800 million Chinese Web pages and corresponding Web spam training sets. This data set was used for evaluating performance of the proposed spam detection framework.

AIRWeb'08, April 22, 2008 Beijing, China.

[1] Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60621062, 60503064, 60736044) and National 863 High Technology Project (2006AA01Z141)

The rest of the paper is organized as follows: Section 2 gives a brief review of related work in Web spam detection. Section 3 analyzes the differences in user-visiting patterns between Web spam and ordinary pages. The spam detection framework based on behavior analysis and Bayesian Learning is proposed in Section 4 and experimental results are presented in Section 5 by performance evaluation on large scale Web access logs. In the end, there is the conclusion and a discussion of future work.

2. RELATED WORK

2.1 Web spamming techniques

According to Gyongyi's Web spamming taxonomy proposed in [3], spamming techniques are grouped into two categories: term spamming and link spamming.

Term spamming refers to techniques that tailor the contents of special HTML text fields in order to make spam pages relevant for some queries [3]. HTML fields which are often adopted by spamming techniques include page titles, keywords in the Meta field, URLs and hyper-link anchors. Hot search keywords are listed (sometimes repeatedly) in these fields to get high rankings by cheating search engines' content relevance calculation algorithms.

Link spammers create hyper-link structures to optimize their scores in hyper-link structure analysis algorithms. Link analysis algorithms such as PageRank [5] and HITS [6] are usually adopted to evaluate the importance of Web pages. Link farms, honey pots and spam link exchange are all means of trickily manipulating the link graph to confuse these algorithms.

After the Gyongyi's spam taxonomy was proposed in 2005, many more spam types appears on the Web and it is difficult to group some of them into the proposed categories. Spam pages' content crawled by Web search spiders may differ from what users would see by cloaking techniques [7]. Browsers may be redirected to visit third-party spam domains when users want to browse "normal" pages [8]. JavaScript, Cascading Style Sheet (CSS) or even Flash movies are currently being adopted by spammers (See Figure 1).

Figure 1. A Web spam page which adopts JavaScript techniques to hide ads. The cell phone ring tone download ads are hidden in the JavaScript http://www.xinw.cn/10086/go1.js. Left: HTML texts of the page; Right: appearance of the page.

2.2 Web spam detection algorithms

Once a new type of Web spam appears on the Web, an anti-spam technique will be developed to identify it. Then new Web spam techniques will be implemented to confuse that technique and so on. In order to combat Web spam and improve search user experience, search engines and Web search researchers have developed lots of methods to detect Web spam pages.

Sometimes spamming techniques can be detected by analysis into statistical content-based attributes of page contents, such as Fetterly et al [9] and Ntoulas et al [10]. Most Web spam identification efforts were focused on hyper-link structure analysis. Davison [11] and Amitay et al [12] are among the earliest ones who study in link spam. Gyongyi et al. [13] proposed the TrustRank algorithm to separate reputable pages from spam. His work was followed by much effort in spam page link analysis such as Anti-Trust Rank [14] and Truncated PageRank [15]. Learning-based methods are also adopted to combine hyperlink features to get better detection performance [16]. Besides that, Wu and Davison proposed in [7] an anti-cloaking method by crawl and compare different copies of a Web page. Wang et al. [8] propose to identify redirection spam pages by connecting spammers and advertisers through redirection analysis. Svore et al [17] adopt query-dependent features to improve spam detection performance.

These anti-spam techniques can detect specific types of Web spam and most of them can achieve good identification performance. However, because there are always new types of spamming techniques, Web spam can still be found in a search engine's result list, sometimes at high ranking positions. There are two major problems with these spam detection methods:

1. The "multi-type problem": most state-of-the-art anti-spam techniques are designed to deal with single type of Web spam, therefore, it makes search engine's anti-spam process a much complicated one because it has to identify all current types of spam.
2. The "prediction problem": it still remains a problem that a Web spam type is difficult to be identified at an early stage before it brought too much discomfort to search users.

With user behavior analysis, we tried to solve the two problems and improve search engine's spam identification performance.

3. USER BEHAVIOR FEATURES OF WEB SPAM PAGES

For WWW information providers, understanding user-visiting patterns is essential for effective Web site designing. Therefore, lots of commercial Web sites collect user access log of pages through software such as on-line service servers or Web browsers.

In order to analyze the behavior pattern of Web users, we collected Web access log from July 1st, 2007 to August 26th, 2007 with the help of a commercial search engine company. No private information is included in these access logs but user sessions can be identified by different session IDs. The access log contains over 2.74 billion user clicks in 800 million Web pages and 22.1 million user sessions during 57 days. Information recorded in the logs is shown in Table 1. During the time period in which the access log was collected, we had three assessors examine the search result lists of the 1000 most frequently asked queries in Sogou search engine (http://www.sogou.com/). By this means, 802 spam sites were identified and used in feature selection process as training set.

With these access logs and spam training set, we were able to look into the different behavior patterns between ordinary and spam pages in order to better understand the perceptual and cognitive factors underlying Web user behaviors. Based on analysis into these differences, we propose three user behavior features to separate Web spam from ordinary pages.

Table 1. Information recorded in Web access logs

Name	Description
Session ID	A random assigned ID for each user session
Source URL	URL of the page which the user is visiting
Destination URL	URL of the page which the user navigates to
Stay time	Stay time of the source page (in seconds)

3.1 Search Engine Oriented Visiting Rate

People visit Web pages through various ways: they may get one nice Web site's recommendation from friends or trustable ads; they may revisit valuable pages in their browsers' bookmark or history lists; they may also follow certain Web pages' out-links according to their interest.

Spam pages try to attract Web user's attention but its content is not valuable for most search users. Therefore, few people will get a spam page's recommendation from a friend, or save it in their bookmark lists, or visit it by following a non-spam pages' hyperlinks. The main or only way for most Web spam pages to be visited is through search result lists. However, for ordinary pages, if it contains some useful information, there are other ways (other person's or Web page's recommendation) besides search result list.

We define the Search Engine Oriented Visiting rate (*SEOV* rate) of a certain page *p* as:

$$SEOV(p) = \frac{\#(Search\ engine\ oriented\ visits\ of\ \ p)}{\#(Visits\ of\ \ p)} \quad (1)$$

It is seldom for Web spam pages to be visited except through search result lists; but ordinary pages may be visited by other means. Therefore, the *SEOV* values of Web spam pages should be higher than ordinary pages. Our statistical result in Figure 2 validates this assumption.

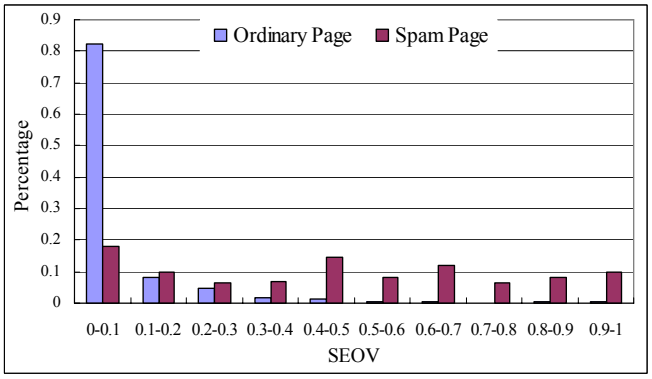


Figure 2. Search Engine Oriented Visiting (SEOV) rate distribution of ordinary pages and Web spam pages. (Category axis: SEOV value interval; Value axis: percentage of pages with corresponding SEOV values.)

In Figure 2, 82% ordinary pages get less than 10% of their visits from search engines; while almost 60% Web spam pages receive over 40% navigation from search result lists. Furthermore, there is less than 1% ordinary Web pages in our corpus with *SEOV* values that are over 0.7; while over 20% spam pages' *SEOV* values are over 0.7. Therefore, we can see that most Web spam pages' *SEOV* value is higher than ordinary pages because search engine is the target of Web spamming and sometimes the only way in which spam can be visited.

3.2 Source Page Rate

Once a hyperlink is clicked, URLs of both the source page and the destination page are recorded in the Web access log. For each page, it may appear either as a source page or as a destination page. However, we found that Web spam are rarely recorded as source pages. Although spam pages may contain hundreds or even thousands of hyperlinks such as the page shown in Figure 3, the hyperlinks on them are hardly clicked by most Web users.

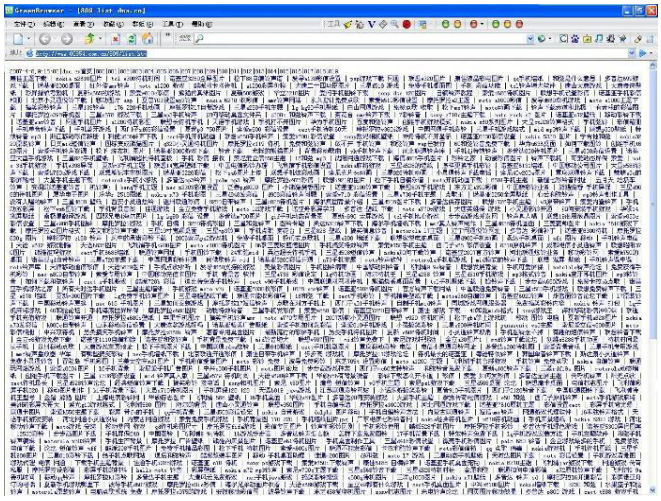


Figure 3. A Web spam page which contains lots of hot keywords to attract search users. All keywords are anchor texts whose hyperlinks link to a same advertisement for Value-added telecom services.

We can define the Source Page (SP) rate of a given Web page *p* as the number of *p*'s appearances as a source page divided by the number of *p*'s appearances in the Web access logs. That is:

$$SP(p) = \frac{\#(p\ appears\ as\ the\ source\ page)}{\#(p\ appears\ in\ the\ Web\ access\ logs)} \quad (2)$$

Experimental results in Figure 4 show the *SP* distribution of ordinary pages and spam pages. We can see that most ordinary pages' *SP* values are larger than those of spam pages. Almost half of the spam pages in the training set rarely appear as the source page (*SP* < 0.05). Only 7.7% spam pages' *SP* rates are over 0.40, while for ordinary pages the percentage is over 53%.

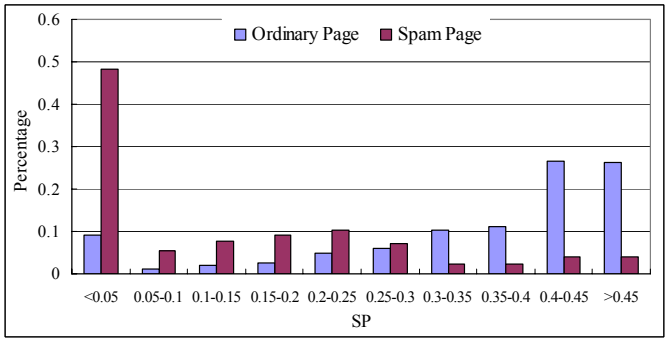


Figure 4. Source Page (SP) rate distribution of ordinary pages and Web spam pages. (Category axis: SP value interval; Value axis: percentage of pages with corresponding SP values.)

The differences in *SP* value distribution can be explained by the fact that spam pages are usually designed to show users misleading advertisements or low-quality services at the first look. Therefore,

most Web users will not click the hyperlinks on spam pages as soon as they notice the spamming activities. Lots of spam pages hardly appear as source pages because when users visit these pages via hyperlinks, they will end their navigation and follow hyperlinks on other pages.

3.3 Short-time Navigation Rate

User attention is the one of the most important resources for WWW information providers. Improving the numbers of user visits and page visits is essential for most commercial Web sites. Therefore, ordinary Web site owners always want to keep users navigating in their sites as long as possible.

However, things are different for Web spammers. Instead of retaining users in their web sites, spammers' major purpose to construct Web spam sites is to guide users to advertisements or services they wouldn't like to see. They don't expect Web users to navigate inside their Web sites; therefore, when users visit any page in a spam site, advertisements or services are usually shown to them at their first look. Meanwhile, this spamming activity causes most Web users to end their navigation in spam sites at once because they don't expect to see such contents. Therefore, we can assume that most Web users would not visit a lot of pages inside spam Web sites. We define the Short-time Navigation rate (*SN* rate) of a web site to describe this assumption. *SN* rate of a given Web site *s* is defined as:

$$SN(s) = \frac{\#(Sessions\ in\ which\ users\ visit\ less\ than\ N\ \ pages\ in\ s)}{\#(Sessions\ in\ which\ users\ visit\ s)} \quad (3)$$

Differently with *SEOV* and *SP*, *SN* is a site-based feature to identify Web spamming techniques. The threshold *N* in its definition is set to 3 in our researches.

Most Web users will not continue their visits inside a spam site, but many of them may visit a number of pages in ordinary Web sites because these sites are designed to keep users staying inside them. Therefore, *SN* rates of Web spam sites should be much higher than those of ordinary Web sites. Our statistical result in Figure 5 validates this assumption.

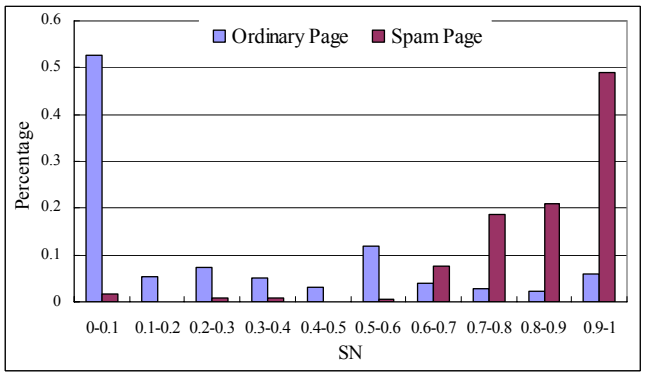


Figure 5. Short-time Navigation (*SN*) rate distribution of ordinary Web sites and Web spam sites. (Category axis: *SN* value interval; Value axis: percentage of Web sites with corresponding *SN* values.)

In Figure 5, 53% ordinary Web sites' *SN* value is less than 0.1, which means over 90% of their visiting sessions contains more than 2 page visits (as mentioned before, *N* is set to 3 in our *SN* definition). However, only 14% of the Web spam sites have *SN* values which are less than 0.1. Meanwhile, 35% Web spam sites have over 0.80 opportunities that users visit only 1 or 2 pages inside them before stop navigation. Therefore, we can see that most Web spam pages' *SN* value is higher than ordinary pages because they cannot and have no intention to keep users staying in their sites.

4. USER BEHAVIOR BASED SPAM DETECTION ALGORITHM

In order to combine the user-behavior features mentioned in Section 3, we try to use a learning-based mechanism to finish the Web spam detection task.

Web spam detection has been viewed as a classification problem in previous works such as [17]. Web spam page classification shares a similar difficulty with the Web page classification problem described by Yu, Han, and Chang [21] in the lack of negative examples. Positive examples (Web spam pages) can be annotated by a number of assessors using techniques such as pooling [22] and our algorithm doesn't require specific spamming types. However, there are so many ordinary pages and a uniform sampling without bias is almost impossible because uniform sampling is regarded as a challenge for Web researchers [4].

Several learning mechanisms were proposed to accomplish the task of Web page classification based on unlabelled data and a number of positive examples. Techniques such as PEBL learning framework [21], semi-supervised learning [23], single-class learning [24] and one class SVM (OSVM) [25] have been adopted to solve the problem. Unlike these algorithms, our anti-spam approach is based on naïve Bayesian Learning method [26] which is believed to be both effective and efficient for low dimensional instance spaces.

We adopt Bayesian learning because it is among the most practical and effective approaches for the problem of learning to classify documents or Web pages. It can also provide explicit probabilities of whether a Web page is a spam page, which can be potentially adopted in result ranking of search engines.

For the problem of Web spam classification, we consider two cases, i.e., the case where classification is based on only one feature and the case where multiple features are involved.

Case 1: Single feature analysis. If we adopt only one user-behavior feature *A*, the probability of a web page *p* with feature *A* being a Web spam can be denoted by

$$P(p \in Spam \mid p\ has\ \ feature\ A) \quad (4)$$

We can use Bayes theorem to rewrite this expression as

$$P(p \in Spam \mid p\ has\ \ feature\ A) = \frac{P(p\ has\ \ feature\ A \mid p \in Spam)}{P(p\ has\ \ feature\ A)} \times P(p \in Spam) \quad (5)$$

In Equation (5), $P(p \in Spam)$ is the proportion of spam pages in the whole page set. This proportion is difficult to be estimated in many cases, including our problem of Web spam page classification. However, if we just compare the values of $P(p \in Spam \mid p\ has\ \ feature\ A)$ in a given Web page corpus, $P(p \in Spam)$ can be regarded as a constant value and it wouldn't affect the comparative results. So in a fixed corpus, we can rewrite equation (5) as:

$$P(p \in Spam \mid p\ has\ \ feature\ A) \propto \frac{P(p\ has\ \ feature\ A \mid p \in Spam)}{P(p\ has\ \ feature\ A)} \quad (6)$$

Now consider the terms in Equation (6), $P(p\ has\ feature A \mid p \in Spam)$ can be estimated using the proportion of *A*-featured pages in the

Web spam page set. While $P(p\ has\ feature\ A)$ equals the proportion of the pages with feature A in a given corpus. Here we obtain:

$$\frac{P(p\ has\ feature\ A \mid p \in Spam)}{P(p\ has\ feature\ A)} = \frac{\#(p\ has\ feature\ A \cap p \in Spam)}{\#(Spam)} \Big/ \frac{\#(p\ has\ feature\ A)}{\#(CORPUS)} \quad (7)$$

If the sampling of Web spam page can be regarded as a uniform process approximately, we can rewrite the numerator of (7) as:

$$\frac{\#(p\ has\ feature\ A \cap p \in Spam)}{\#(Spam)} = \frac{\#(p\ has\ feature\ A \cap p \in Spam\ sample\ set)}{\#(Spam\ sample\ set)} \quad (8)$$

Substituting expressions (7) and (8) into (6), we obtain:

$$P(p \in Spam \mid p\ has\ feature\ A) \propto \frac{\#(p\ has\ feature\ A \cap p \in Spam\ sample\ set)}{\#(Spam\ sample\ set)} \Big/ \frac{\#(p\ has\ feature\ A)}{\#(CORPUS)} \quad (9)$$

Since all terms in (9) can be obtained by statistical analysis on a Web page corpus, we can calculate the probability of being a Web spam for each page according to this equation.

Case 2: Multiple feature analysis. If we use more than one feature to identify Web spam pages, naïve Bayes theorem assumes that the following equation holds:

$$P(p\ has\ feature\ A_1, A_2, ..., A_n \mid p \in Spam) = \prod_{i=1}^{n} P(p\ has\ feature\ A_i \mid p \in Spam) \quad (10)$$

For the problem of page classification with user-behavior features, we further found that the following equation also approximately holds according to Table 2.

$$P(p\ has\ feature\ A_1, A_2, ..., A_n) = \prod_{i=1}^{n} P(p\ has\ feature\ A_i) \quad (11)$$

Table 2 Correlation values between user-behavior features of Web pages

	SEOV	***SP***	***SN***
SEOV	1.0000	0.1981	0.1780
SP	0.1981	1.0000	0.0460
SN	0.1780	0.0460	1.0000

The correlation values in Table 2 show that these three features are approximately independent of one another because the values are all close to zero. This may be explained by the fact that these features are obtained from different information sources and thus have little chance affecting one another. This means for the features in Table 2, the attribute values adopted in the Web spam page classification process are independent as well as conditionally independent given the target value.

The following equations hold approximately for the Web spam page classification task according to naïve Bayes assumption and our statistical analysis:

$$\begin{aligned} &P(p \in Spam \mid p\ has\ feature\ A_1, A_2, ..., A_n) \\ &= \frac{P(p\ has\ feature\ A_1, A_2, ..., A_n \mid p \in Spam) P(p \in Spam)}{P(p\ has\ feature\ A_1, A_2, ..., A_n)} \\ &\approx \prod_{i=1}^{n} \frac{P(p\ has\ feature\ A_i \mid p \in Spam) P(p \in Spam)}{P(p\ has\ feature\ A_i)} \\ &= \prod_{i=1}^{n} P(p \in Spam \mid p\ has\ feature\ A_i) \end{aligned} \quad (12)$$

If we substitute (9) into (12), we can get the following equation which is fit for multi-feature cases:

$$P(p \in Spam \mid p\ has\ feature\ A_1, A_2, ..., A_n) \propto \prod_{i=1}^{n} \left(\frac{\#(p\ has\ feature\ A_i \cap p \in Spam\ sample\ set)}{\#(Spam\ sample\ set)} \Big/ \frac{\#(p\ has\ feature\ A_i)}{\#(CORPUS)} \right) \quad (13)$$

According to this equation, the probability of a web page being a Web spam page can be calculated with information from the Web corpus and its corresponding spam page sample set. Therefore it is possible for us to use the following algorithm to accomplish the spam identification task.

Algorithm 1. Web spam detection with user behavior analysis

1. Collect Web access log (with information shown in Table1) and construct access log corpus *S*;
2. Calculate *SEOV* and *SP* scores according to Equation (1) and (2) for each Web page in *S*;
3. Calculate *SEOV* and *SP* scores for each Web site in *S* by averaging scores of all pages in the site;
4. Calculate *SN* score for each Web site in *S* according to Equation (3);
5. Calculate *P(Spam | SEOV, SP, SN)* according to Equation (9) for each Web page in *S*.

After performing **Algorithm 1** on *S*, we will get a spam probability score for each Web page in *S*.

5. EXPERIMENTS AND DISCUSSIONS

5.1 Detection of Various Kinds of Spam Pages

After Bayesian learning on the training set, we construct a classifier which can assign possibilities of being spam to Web pages based on user behavior analysis. Then we randomly sampled 1564 Web sites from the access log (about 1/1000 of all Web sites covered in the corpus) and have three assessors to annotate these sites as "spam", "non-spam" or "cannot tell". The results show that 345 sites are spam, 1060 are non-spam and assessors "cannot tell" 159 sites whether are spam or not.

After the annotation, we choose ROC curves and corresponding AUC values to evaluate the performance of our spam detection algorithm. It is a useful technique for organizing classifiers and visualizing their performance and it is also adopted by several other Web spam detection researches such as [17] and [27]. ROC curves of the spam detection algorithm are shown in Figure 6.

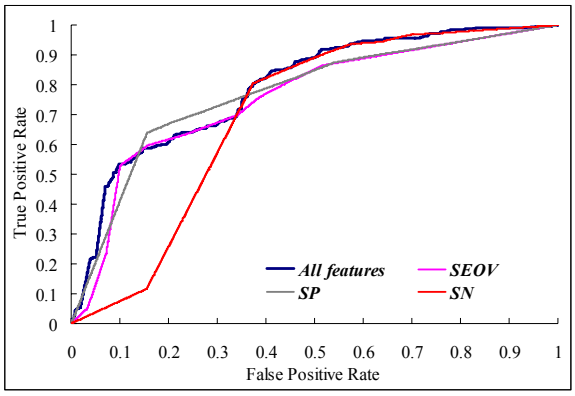


Figure 6. ROC curves on test sets using Bayesian learning to combine *SEOV*, *SP* and *SN* features, compared with the curves on test sets using a certain user-behavior feature only.

We can see in this Figure that *SP* and *SEOV* are more effective than *SN* in detecting Web spam. However, the learning algorithm proposed in Section 4 combines all features and gains better performance than any of the three features.

The AUC value for the algorithm's ROC curve is 0.7926, which means our detection algorithm has 79.26% chances to rank a Web spam higher than a non-spam in the spam-possibility result list. It is not as high as the AUC scores obtained by algorithms proposed in previous works such as the ones in Web Spam Challenge (http://webspam.lip6.fr/). However, we believe that our experiment settings are more close to practical Web search applications by using large scale datasets (over 10 times larger than the one adopted by Web Spam Challenge) and more efficient classifiers (only 3 features are involved and computation complexity is O(N)). Besides, we found that the user-behavior based algorithm is able to identify various kinds of spam pages. According to the experimental results in Table 3, term-based, link-based and other kinds of spamming techniques can all be detected by the algorithm.

Table 3. Page types of 300 possible spam pages identified by our spam detection method

Page Type	Percentage
Non-spam pages	6.00%
Web spam pages (Term spamming)	21.67%
Web spam pages (Link spamming)	23.33%
Web spam pages (Other spamming)	10.67%
Pages that cannot be accessed	38.33%

In Table 3, 300 pages which are identified as spam by our algorithm are annotated with their page types. These pages are the top-ranked ones in the possible spam list ranked by spam probabilities. Firstly, we found that most of the identified pages are spam pages and about 6% of these pages are not spam pages. However, further analysis into these non-spam pages shows that they are mostly low-quality pages that adopt some kind of SEO techniques to attract users. Secondly, there are also a number of pages which cannot be accessed at the time of assessment. We believe that most of these pages are previously spam because spam pages usually change their URL to bypass search engines' spam list. Meanwhile, ordinary pages wouldn't change their domain name because that hurts their rankings in search engines. Finally, we can see that both term-based and link-based spamming techniques can be identified by our algorithm. We adopt user behavior features to detect Web spam, which makes it possible to identify Web spam independent of spamming technique types. This can be regarded as a possible solution to the "multi-type problem" proposed in Section 2.2.

5.2 Detection of Newly-appeared Spam Pages

We mentioned the "prediction problem" in Section 2.2 and regarded it as one of the most challenging problems for state-of-the-art anti-spam techniques. We found that our algorithm can identify various kinds of Web spam pages and we still want these spam pages to be identified as soon as possible. Therefore, we designed the following experiments to see whether our algorithm can identify Web spam more timely than the spam detection methods adopted by commercial search engines.

We divided the Web access data mentioned in Section 3 into two parts. The first part includes access log in July 2007 and are adopted to train a spam classifier using our detection algorithm. Then the spam classifier ranked Web pages by their possibilities of being spam. We choose the top-ranked 1200 pages as possible spam pages and see what will happen to them in August.

Figure 7 compares search engine oriented user-visiting between the top-ranked possible spam pages and 44000 random-sampled ordinary Web pages. Four widely-used Chinese search engines (Baidu, Yahoo! China, Google China and Sogou) were used to test the search oriented user-visiting.

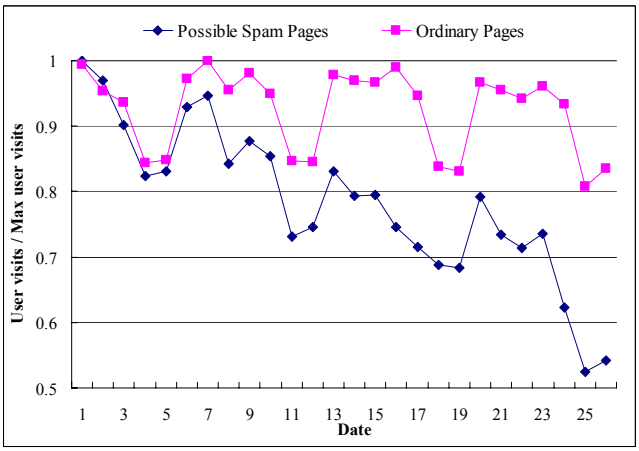


Figure 7. Search engine oriented visiting of possible spam pages and ordinary pages. (Category axis: date in August, 2007; Value axis: total number of search engine oriented visiting of the page set in a certain day divided by the maximum number of search-oriented visiting during a single day in these days.)

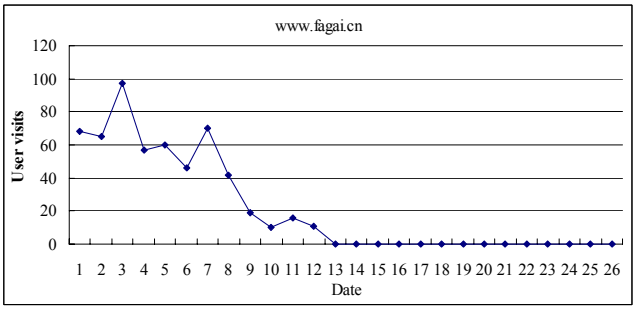
In Figure 7, we can see that the search-oriented visiting of ordinary pages almost retains a same number during these days (2007/08/01 – 2007/08/26). The user visits during weekdays is about 15% higher than that at weekends because people use Web search engines to obtain information more frequently during weekdays. User-visiting number of the day with the fewest user-visits (in August 25th) is about 80% of that during the day with the most visits (in August 1st and 7th). Ordinary page contains a relatively fixed amount of information. For a single Web page, its user-visits may vary substantially (for example, some political event happens and a certain statesman's page may be viewed many times) but as a whole it is not likely to change significantly within several days.

However, things are much different for those possible spam pages. Number of user visits from search engines went down with

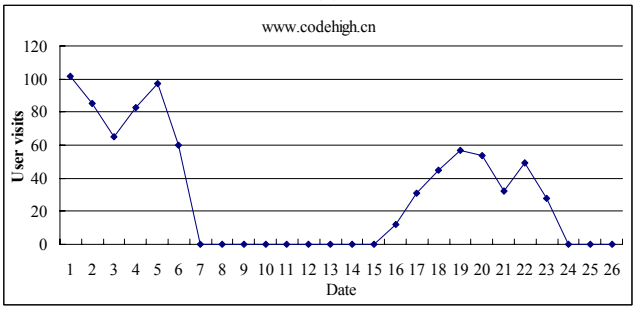
time apparently during these days. Number of user visits in the last day is only half of the visiting number of the first day. 15 out of 26 days have less than 80% user visits of the first day.

This phenomenon can be explained as follows: a number of pages in the possible spam list can be regarded newly-appeared Web spam pages. They receive plenty of user visits from search engines at the beginning of this month because they are not identified by search engines yet. As time passes, search engines detected them with anti-spam techniques and some of them are reduced from search result lists. Therefore, these pages receive far less user visits at the end of August. A large number of these pages are Web spam pages but search engine didn't identify them at the end of July as our algorithm did.

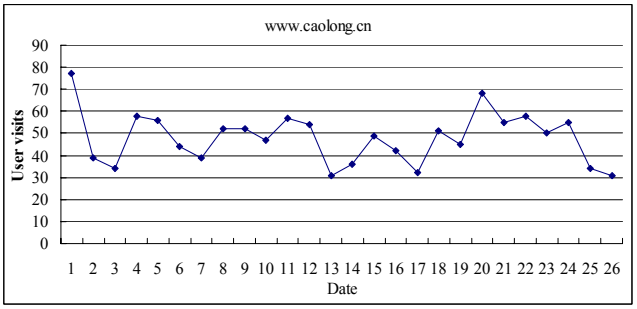
We further look into several identified spam sites in the possible spam set and see how their user-visiting number varies with time. Examples of these Web spam sites are shown in Figure 8.



(a)



(b)



(c)

Figure 8. User visits of three spam sites identified by our spam detection method. (Category axis: date in August, Value axis: search engine oriented user visits of corresponding site.)

Figure 8 shows three Web spam sites which are identified by our spam detection method using Web access log in July. The site http://www.fagai.cn/ (shown in Figure 8(a)) was also identified by search engines in early August and its search-oriented user visits dropped down to nothing at the middle of August. http://www.codehigh.cn (shown in 8(b)) is identified by a certain search engine on August 5th or 6th and then its user visits dropped significantly. However, in August 15th it might be crawled by another search engine and search-oriented visiting went up to about 60 times per day. At last, it was identified as spam again and user-visiting number returns to nothing. http://www.caolong.cn/ is a Web spam page which was not recognized by search engines during August so its user visits stayed relatively stable.

The spam page shown in Figure 8(a) and Figure 8(b) are not detected by search engines as timely as our algorithm did. The one shown in Figure 8(c) wasn't even detected at the end of the month. It shows that our method is able to detect newly-appeared Web spam pages and the detection is more timely than the anti-spam techniques adopted by commercial search engines.

6. CONCLUSIONS AND FUTURE WORK

Most spam detection approaches focus on predefined types of spam pages using content or hyperlink analysis. Different from this traditional method, we propose a user-behavior oriented Web spam detection algorithm. This algorithm analyzes large-scale Web access logs, and exploits the differences between Web spam pages and ordinary pages in user behavior patterns. We combine machine learning techniques and descriptive analysis on user-behavior features of Web spam pages. By this means, we come to a better and further understanding of the relationship between user visiting pattern and Web page quality.

Currently, the user-behavior oriented approach may not be as effective as state-of-the-art anti-spam algorithms in identifying certain types of Web spam. However, with the help of Web user behavior, it can detect various kinds of spam pages no matter what spamming techniques they adopt. Newly-appeared spam can also be identified as soon as a number of users are bothered by them. This method may not replace existing anti-spam algorithms but it can help search engines to find out the most bothersome spam types and be aware of newly-appeared spam techniques.

In the near future, we hope to extend this framework to embody page content and hyperlink features. We also plan to work on a Web page quality estimation model for Web search tools based on our findings in this paper.

7. REFERENCES

[1] CNNIC (China Internet Network Information Center), the 16th report in development of Internet in China. Online at http://www.cnnic.net.cn/uploadfiles/pdf/2005/7/20/210342.pdf.

[2] Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. 1999. Analysis of a very large web search engine query log. SIGIR Forum 33, 1 (Sep. 1999), 6-12.

[3] Gyongyi, Z. and Garcia-Molina, H. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[4] Henzinger, M.R., Motwani, R., Silverstein, C. 2003. Challenges in Web Search Engines. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (2003) 1573-1579.

[5] Brin, S. and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the Seventh international Conference on World Wide Web 7 (Brisbane, Australia). 107-117.

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

[7] Wu, B. and Davison, B. Cloaking and redirection: a preliminary study. In First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '05), May 2005.

[8] Wang, Y., Ma, M., Niu, Y., and Chen, H. Spam double-funnel: Connecting web spammers with advertisers. In Proc. of the 16th International Conference World Wide Web (WWW), May 2007.

[9] Fetterly, D., Manasse, M. and Najork, M. Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In S. Amer-Yahia and L. Gravano, editors, WebDB, pages 1–6, 2004.

[10] Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. 2006. Detecting spam web pages through content analysis. In Proceedings of the 15th International Conference on World Wide Web (Edinburgh, Scotland, May 23 - 26, 2006). WWW '06. ACM Press, New York, NY, 83-92.

[11] Davison B. Recognizing nepotistic links on the Web. In Artificial Intelligence for Web Search, pages 23--28. AAAI Press, July 2000. Presented at the AAAI-2000 workshop on Artificial Intelligence for Web Search, Technical Report WS-00-01.

[12] Amitay, E., Carmel, D., Darlow, A., Lempel, R., and Soffer, A. 2003. The connectivity sonar: detecting site functionality by structural patterns. In Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia (Nottingham, UK, August 26 - 30, 2003). HYPERTEXT '03.

[13] Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. 2004. Combating web spam with trustrank. In Proceedings of the Thirtieth international Conference on Very Large Data Bases - Volume 30. 576-587.

[14] Krishnan, V. and Raj, R. Web Spam Detection with Anti-Trust-Rank. In the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), August 2006.

[15] L. Becchetti, C. Castillo1 D. Donato1, S. Leonardi, and R. Baeza-Yates. Using Rank Propagation and Probabilistic Counting for Link Based Spam Detection. In Proc. of WebKDD'06, August 2006.

[16] Geng, G., Wang, C., Li, Q., Xu, L., and Jin, X. 2007. Boosting the Performance of Web Spam Detection with Ensemble Under-Sampling Classification. In Proceedings of the Fourth international Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007) Vol.4 - Volume 04 (August 24 - 27, 2007). FSKD. IEEE Computer Society, Washington, DC, 583-587.

[17] Svore, K., Wu, Q., Burges, C. and Raman, A. Improving Web Spam Classification using Rank-time Features. In Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '07), May 2007.

[18] Sullivan D. 2006. Searches Per Day. Retrieved from search engine watch web site http://searchenginewatch.com/reports/article.php/2156461.

[19] Silverstein, C., Marais, H., Henzinger, M., and Moricz, M. 1999. Analysis of a very large web search engine query log. SIGIR Forum 33, 1 (Sep. 1999), 6-12.

[20] Yu, H., Liu, Y., Zhang, M. and Ma, S. Research in Search Engine User Behavior Based on Log Analysis. Journal of Chinese Information Processing. Vol. 21(1): pp. 109-114, 2007.

[21] Yu, H., Han, J., and Chang, K. C. 2004. PEBL: Web Page Classification without Negative Examples. IEEE Transactions on Knowledge and Data Engineering 16, 1 (Jan. 2004), 70-81.

[22] Voorhees. E. M. 2001. The philosophy of information retrieval evaluation. In Proceedings of the Second Workshop of the Cross-Language Evaluation Forum, (CLEF 2001), pages 355-370.

[23] Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning. 39(2-3): 103-134.

[24] Denis, F. PAC Learning from Positive Statistical Queries (pp.112-126). Proceedings of the 9th international Conference on Algorithmic Learning theory. Lecture Notes In Computer Science, vol. 1501. London: Springer-Verlag, 1998.

[25] Manevitz, L. M. & Yousef, M. One-class SVMs for document classification. Machine Learning. Res. 2: 139-154.

[26] Mitchell, T. Chapter 6: Bayesian Learning, in Mitchell, T., Machine Learning, McGraw-Hill Education, 1997.

[27] Web Spam Challenge Website: http://webspam.lip6.fr/