User Behavior Oriented Web Spam Detection[1]

Yiqun Liu, 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 recently-appeared spam. With user behavior analyses into Web access logs, we propose a spam page detection algorithm based on Bayes learning. Preliminary experiments on Web access data collected by a commercial Web site (containing over 2.74 billion user clicks in 2 months) show the effectiveness of the proposed detection framework and algorithm.

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. According to [1], most search users only view the first few URLs in result lists, so internet service providers want their pages to be ranked as high as possible by search engines to capture more user attention. Web spam can be defined as any attempt to get "an unjustifiably favorable relevance or importance score for some Web page, considering the page's true value" [2]. Because Web spam leads to obstacles to users' information acquisition process, spam detection is treated as a major challenge for search engines.

Currently, anti-spam techniques usually make use of Web page's content [3] or hyper-link features [4] to construct classifier and identify spam pages. After a certain kind of Web spam appears in search result lists, engineers examine the characteristics of this spam type and design the specific strategies to identify it. However, once a kind of spam is detected and banned, the spammers will turn to develop new Web spam instantly. With this method, anti-spam techniques can only identify Web spam which has already caused severe loss and drawn search engineers' attention.

In contrast to the prevailing approaches, we propose a different anti-spam framework in which spam sites are identified because of their deceitful motivation instead of their content/hyper-link appearance. We introduce three features developed from user behavior pattern analyses and design a learning-based approach to combine these behavior features to identify Web spam pages.

2. USER-BASED SPAM DETECTION

In order to analyze into the behavior pattern of Web users, we collected Web access log from July 1st, 2007 to August 26th, 2007 with the help of Sohu.com using browser toolbars (http://tb.sogou.com/). No private information is included in these access logs but user sessions can be identified by different session IDs. Source and Destination URLs of user clicks and user's stay time in a certain page are also recorded. The access log contains over 2.74 billion user clicks on 800 million Web pages and 22.1 million user sessions during 57 days.

We also construct Web spam training and test sets to verify the effectiveness of the proposed algorithm. 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. A different set of 1149 Web sites were random selected from the sites whose access behavior was recorded. These sites were annotated as spam or not and used for performance evaluation.

2.1 User Behavior Feature of Web Spam Page

Based on analysis into the different user behavior patterns between Web spam pages and ordinary pages, we propose three features to identify spam pages.

Firstly, spam pages try to attract Web user's attention but its content is not as valuable as it appears. Therefore, Web spam page receives most of its user visiting from search engines instead of from non-spam pages or bookmark lists. We define the Search Engine Oriented Visiting rate (*SEOV* rate) of a certain page *p* as:

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

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.

Secondly, user attention is one of the most important resources for WWW information providers. Ordinary Web site owners always want to keep users navigating in their sites as long as possible. However, spammers' major purpose to construct spam sites is to guide users to advertisements or services they wouldn't like to see. Therefore, most Web users tend to end their navigation in spam sites as soon as they noticed the spamming activities.

We can define two behavior features to describe this different user visiting pattern between ordinary and spam pages. The Start Point Visiting rate (*SP* rate) feature describes how many clicks are performed on a certain page *p* and it can be defined as:

$$SP(p) = \frac{\#(user\ clicks\ a\ hyperlink\ on\ p\ while\ visiting\ \ p)}{\#(Visits\ of\ \ p)} \qquad (2)$$

The Short-term Navigation rate (*SN* rate) shows how many pages of a site *s* will be visited once user visit *s* and can be defined as:

Copyright is held by the author/owner(s).
WWW 2008, April 21?5, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

[1] Supported by the Chinese National 973 Plan (2004CB318108), Natural Science Foundation (60621062, 60503064, 60736044) and National 863 High Technology Project (2006AA01Z141)

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

In order to validate the effectiveness of the three proposed features, we compared the differences in feature distribution between ordinary page and spam pages. For instance, statistic of the SEOV feature is shown in Figure 1.

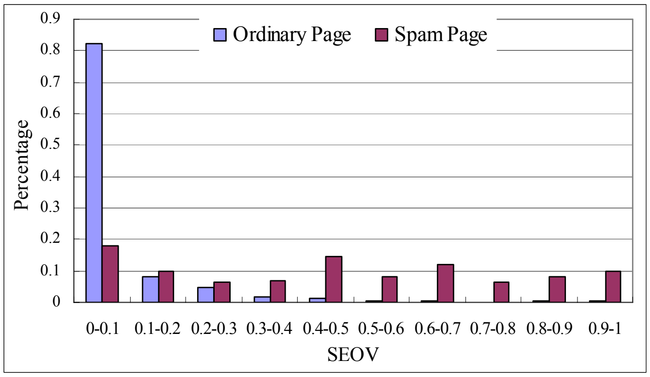

Figure 1. Search Engine Oriented Visiting (SEOV) rate distribution of ordinary and Web spam pages.

In Figure 1, over 80% ordinary pages get less than 10% of their visiting from search engines; while almost 50% Web spam pages receive over half of their navigation from search result lists. 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. The *SP* and *SN* features can also separate Web spam from ordinary pages. According to our statistics, for 48% spam pages, almost none (less than 5%) users click any hyperlink on them ($SP < 0.05$); while 63% ordinary Web pages receive clicks from over 30% of their readers ($SP > 0.3$). For the *SN* feature with N=3, 53% ordinary pages have almost none (less than 10%) short navigations ($SN < 0.1$); meanwhile 96% Web spam's *SN* value is over 0.6.

2.2 Learning-based Spam Detection Algorithm

User-behavior features mentioned in Section 2.1 can be used to identify Web spam pages. In order to combine these features, we try to use naïve Bayes learning method which is believed to be both effective and efficient for low dimensional instance spaces.

$$\begin{aligned}&P(p \in Spam \mid p\ has\ \ feature\ A_1, A_2, ..., A_n) \qquad (4)\\ &= \prod_{i=1}^{n} \left(\frac{\#(p\ has\ \ feature\ A_i \cap p \in Spam\ training\ set)}{\#(Spam\ raining\ set)} \Big/ \frac{\#(p\ has\ \ feature\ A_i)}{\#(Ordinary\ page)} \right)\end{aligned}$$

With Bayes assumption and features' independent assumption, 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 according to Equation (4).

3. EXPERIMENTAL RESULTS

We choose ROC curves and corresponding AUC values to evaluate the performance of our spam detection algorithm because it is a useful technique for organizing classifiers and visualizing their performance.

In Figure 1, *SP* and *SEOV* are more effective than the *SN* feature in detecting Web spam. It also shows that the proposed learning algorithm 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 its spam-possibility result list.

In practical search applications we care more about whether the detected possible spam pages are really Web spam pages. Therefore, we examine the top-ranked Web pages in the spam-possibility list given by our algorithm and analyze their spamming techniques.

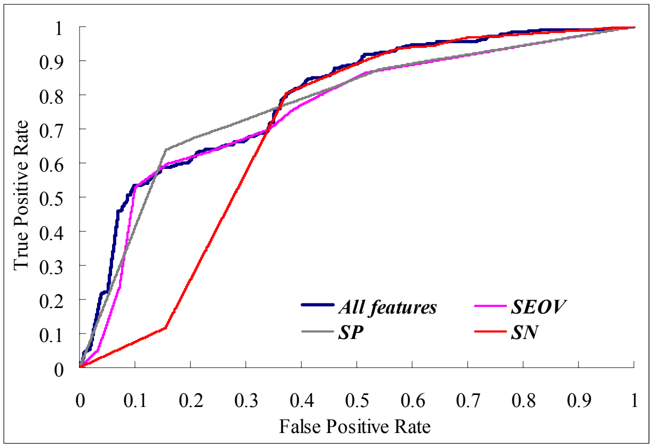

Figure 2. ROC curves on test sets using Bayesian learning to combine *SEOV*, *SP* and *SN* features, compared with the curves on test sets using a certain user-behavior feature only.

Experimental results in Table 1 show that only a few (5.98%) of the top-ranked pages are mis-identified. Analysis into these non-spam pages shows that they are mostly low-quality pages that adopt some kind of SEO techniques to attract users. Reducing them will not cause major loss for most Web users.

Table 1. Types of the top 300 possible spam pages given by user-behavior based spam detection method

Page Type	Percentage
Non-spam pages	6.00%
Web spam pages	55.67%
Pages that cannot be accessed	38.33%

In the top-ranked Web pages of our possible spam list, there are also a number of pages which cannot be accessed at the time of assessment. We believe that most of these pages are 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.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we propose user-behavior-oriented Web spam detection framework, it is a feasible solution to identify Web spam pages effectively and type-independently. In the near future, we hope to extend this framework to embody the state-of-the-art page content and hyperlink features. We also plan to work on a Web page quality estimation model based on our findings in this paper.

REFERENCES

[1] 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.

[2] Zoltán Gyöngyi and Hector Garcia-Molina. Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005.

[3] Ntoulas, A., Najork, M., Manasse, M., and Fetterly, D. Detecting spam Web pages through content analysis. In proceedings of the 15th WWW conference. 83-92.

[4] Gyongyi, Zoltan; Garcia-Molina, Hector; Pedersen, Jan. Combating Web Spam with TrustRank, Proceedings of the 30th International Conference on Very Large Data Bases.