Detecting Link Spam using Temporal Information

Guoyang Shen[1,2*], Bin Gao[1*], Tie-Yan Liu[1], Guang Feng[1,3*], Shiji Song[2], Hang Li[1]

[1]*Microsoft Research Asia*
4F, Sigma Center
No. 49, Zhichun Road
Beijing, 100080, P. R. China
bingao@microsoft.com
tyliu@microsoft.com
hangli@microsoft.com

[2]*National CIMS Engineering Research Center*
Dept. of Automation
Tsinghua University
Beijing, 100084, P. R. China
shengy@mails.tsinghua.edu.cn
shijis@mail.tsinghua.edu.cn

[3]*MSP Laboratory*
Dept. of Electronic Engineering
Tsinghua University
Beijing, 100084, P.R. China
fengg03@mails.tsinghua.edu.cn

Abstract

How to effectively protect against spam on search ranking results is an important issue for contemporary web search engines. This paper addresses the problem of combating one major type of web spam: 'link spam.' Most of the previous work on anti link spam managed to make use of one snapshot of web data to detect spam, and thus it did not take advantage of the fact that link spam tends to result in drastic changes of links in a short time period. To overcome the shortcoming, this paper proposes using temporal information on links in detection of link spam, as well as other information. Specifically, it defines temporal features such as In-link Growth Rate (IGR) and In-link Death Rate (IDR) in a spam classification model (i.e., SVM). Experimental results on web domain graph data show that link spam can be successfully detected with the proposed method.

1. Introduction

Using search engines has become a major means for people to find information on the web. When conducting searches, people usually only look at the top ranked pages returned by search engines [10]. Therefore, a page will have more chances to be accessed, if it is ranked higher by web search engines. Driven by commercial, egotistical, or even malicious motivations, some website or page owners attempt to manipulate information on the web in order to deceive search engines and make their sites or pages ranked high by the engines. This is a practice generally called web spamming.

For a search engine, on the other hand, effectively detecting and eliminating web spam becomes an important issue. In the research community, this has been defined as a research issue and several methods have been proposed [1][2][3][4][5][7][8][9][12][13].

Anti-spam is a challenging task, because new spam techniques are being developed continuously while anti-spam methods are usually created only based on those *known* spamming techniques. Although previous work on anti-spam has made certain progress, more studies on the issue are still needed. Particularly, the development of a method that is robust against different spam techniques, even unknown or potential spam techniques, is sorely needed.

In this paper, we address the issue of anti link spam. Contemporary search engines rank web pages by taking into consideration the number of links connected to the pages. A web page to which more links (called in-links) are connected is more likely to be ranked higher. Spammers, therefore, often attempt to manipulate links on the web, for example, add thousands or even millions of links to the pages which they want to promote – a technique referred to as 'link spam.' Previous work on anti link spam usually utilizes one snapshot of web graph data, extracts features from the data, and identifies spam sites by using the features.

We propose using temporal information in detection of link spam. Specifically, we use several snapshots of web graph data in a time period, extract temporal features as well as other features from the data, and construct an SVM (Support Vector Machines) classifier for the task. The features include *In-link Growth Rate (IGR) and In-link Death Rate (IDR).*

The rationale behind our method is that spam pages and common pages have different evolution patterns along time line. For instance, link spam tends to result in drastic changes of links to spam sites in a short time period. No matter what kind of spam tricks are used, this tendency must be reflected in a time series of web data. Our method, therefore, is very effective and robust for link spam detection. Experimental results on real web graph

*This work was conducted when the first, the second and the fourth authors were interns at Microsoft Research Asia.

Proceedings of the Sixth International Conference on Data Mining (ICDM'06)
0-7695-2701-9/06 $20.00 © 2006 IEEE

data have verified the correctness of our claim, as will be seen below.

2. Data set

We collected two snapshots of website graph, dated 2006-02 and 2006-03. A website graph is a graph in which a website forms a node and the hyperlinks between two websites (if there exists) form edges. We used the two website graphs (denoted as 2006-02 and 2006-03) in our experiments throughout the paper.

The statistics of the two website graphs are shown in Table 1.

Table 1. The statistics on the two website graphs.

	2006-02	2006-03	Changed
Average in-degree	17.2865	17.2456	0.0409
# of websites	41,464,052	41,062,569	401,483

Table 2. The differences between the two website graphs.

# of vanished websites	4,753,588
# of new-born websites	4,352,150

From Table 1, we can see that the total numbers of nodes on the 2006-02 graph and the 2006-03 graph only differ slightly. From Table 2, however, we can see that the websites contained in the two graphs actually have changed significantly. This is partly due to our crawling strategy and partly due to the fact that many websites get updated frequently. This indicates that temporal information contained in website graphs is rich and potentially useful for spam detection.

To conduct spam detection experiments, we asked several human annotators to label a randomly selected portion of websites as spam or non-spam sites. The statistics are shown in Table 3.

Table 3. The statistics on the labeled data.

	Number	Proportion
Total	113,756	100%
Spam	12,020	10.57%
Non-spam	101,736	89.43%

3. Temporal features

In this paper, we formalize the problem of link spam detection as that of classification. More precisely, we take each node (website) on the website graph as an instance, use the labeled instances (websites) to train an SVM classifier, and employ the constructed classifier in link spam detection (i.e., identify whether a new instance is a spam site or not). The SVM classifier mainly contains temporal features.

The key issue for our approach, then, is to define the temporal features. We propose three types of features. First, for each node we define features on the basis of the temporal statistics of the node. Next, we introduce features for each node using the temporal statistics of its neighbors. Finally, we define features for each node based on the correlation among the neighbors of the node. For each type of the temporal features, we give justification on the use of it.

We first give explanations on the notations.

$S_{in}(a,t)$ - The set of in-links of website *a* at time *t*.

$S_{out}(a,t)$ - The set of out-links of website *a* at time *t*.

$|S_{in}(a,t)|$ - The number of websites in $S_{in}(a,t)$.

$G(t)$ - The web graph at time *t*.

t_0 - The time when the crawler finished crawling the 2006-02 graph.

t_1 - The time when the crawler finished crawling the 2006-03 graph.

3.1. Temporal features of individual website

3.1.1 In-link Growth Rate (IGR). IGR of a site is defined as the ratio of the increased number of in-links at the site to the number of original in-links (i.e. the number of in-links at the website in the 2006-02 graph).

$$IGR(a) = \frac{|S_{in}(a,t_1)| - |S_{in}(a,t_0) \bigcap S_{in}(a,t_1)|}{|S_{in}(a,t_0)|}$$

The main trick for spammers to conduct link spam is to add in-links to the pages which they want to promote. Therefore, IGR is a good indicator of link spam. To see whether this is the case, we plot the distributions of the in-link growth rate of all the websites in the two website graphs in Figure 1.

The histogram in the figure represents the distribution of IGR for all the websites. The curve in the figure represents the probability of being spam (i.e. the number of labeled spam websites among the number of labeled websites) when IGR is between 0 and 5 (the x-axis). The left vertical axis corresponds to the histogram and the right vertical axis corresponds to the curve.

From Figure 1, we can see that the IGR values of most websites are in the range of [0, 1], in which the probability of spam is very low. Only a small number of websites have IGR values existing in [1, 5], in which the probability of spam is very high. From this observation, we conclude that there is a clear correlation between IGR and probability of spam. Considering the fact that the average probability of spam in the labeled data is 10% (lower bound for spam detection), we can say that IGR is quite a good feature for spam detection. On the other

hand, using this feature alone is not sufficient; there will be a high rate of false positives.

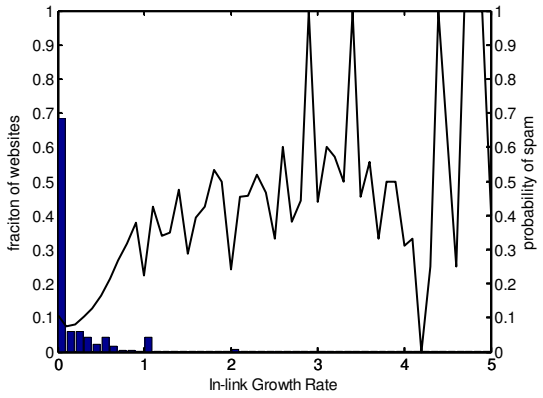

Figure 1. Probability of spam versus IGR.

3.1.2 In-link Death Rate (IDR). IDR of a site is defined as the ratio of the number of dead in-links to the number original in-links (i.e. the number of in-links at the website in the 2006-02 graph).

$$IDR(a) = \frac{|S_{in}(a,t_0)| - |S_{in}(a,t_0) \cap S_{in}(a,t_1)|}{|S_{in}(a,t_0)|}$$

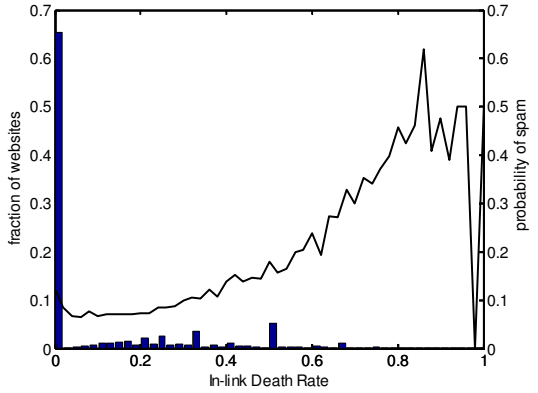

Figure 2. Probability of spam versus IDR.

To avoid being detected by search engines, spammers may frequently change their ways of spamming. This will make the IDR values of the spam sites become higher than those of the normal websites. Again, we use a figure to show the distribution of IDR. From Figure 2, we can see that websites with IDR values higher than 0.4 have a greater probability of being spam.

3.2. Temporal features on the neighbors of a website

3.2.1 Mean of In-links' IGR (IGRMean). IGRMean of a site is defined as average IGR value of its in-link websites.

$$IGRMean(a) = \frac{\sum_{b \in S_{in}(a,t_0)} IGR(b)}{|S_{in}(a,t_0)|}$$

One assumption of BadRank [4] is that bad pages tend to link to bad pages. Following this idea and considering the fact that IGR is a good indicator of spam, we propose utilizing this temporal feature in spam detection. We assume here that a website is likely to be a spam site if the IGR values of its in-link websites are high. Figure 3 shows the justification on using the feature.

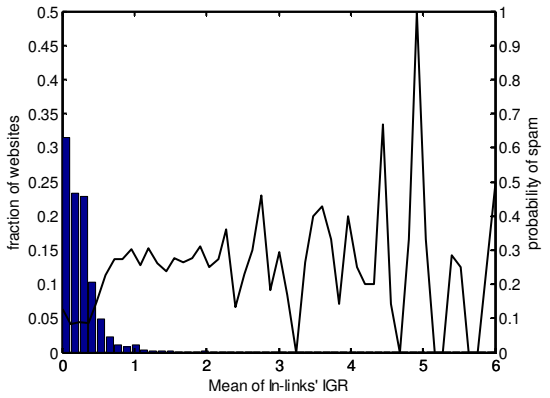

Figure 3. Probability of spam versus IGRMean.

3.2.2 Mean of In-links' IDR (IDRMean). Similar to IGRMean, IDRMean of a site is defined as average IDR value of in-link websites of the site.

$$IDRMean(a) = \frac{\sum_{b \in S_{in}(a,t_0)} IDR(b)}{|S_{in}(a,t_0)|}$$

The motivation of using this feature is similar to that of using IGRMean. We plot the distribution of IDRMean in Figure 4. Comparing Figure 3 and Figure 4, we see that IDRMean can be an even better spam indicator than IGRMean.

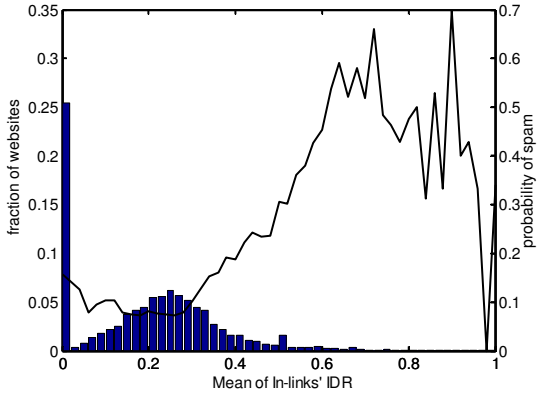

Figure 4. Probability of spam versus IDRMean.

3.2.3 Variance of In-links' IGR (IGRVar). Automated spam sites may have neighbors with very similar

statistical patterns, which normal websites may not have. With this consideration, we calculate the variance of the IGR values of a website's in-link sites, to help spam detection.

$$IGRVar(a) = \sqrt{\frac{\sum_{b \in S_{in}(a,t_0)} \left(IGR(b) - IGRMean(a)\right)^2}{\left|S_{in}(a,t_0)\right|}}$$

From the curve in Figure 5, we can see that most of the websites with high IGRVar values are normal websites and thus IGRVar can be a feature to distinguish spam and non-spam websites.

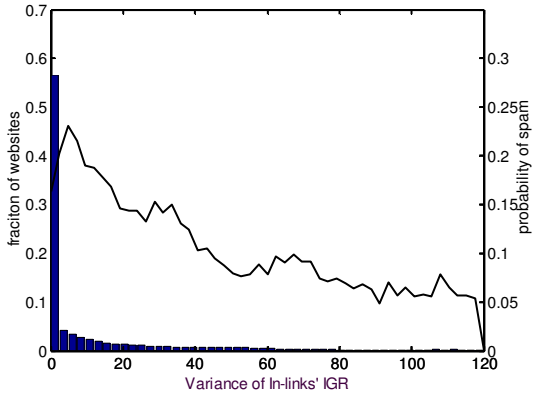

Figure 5. Probability of spam versus IGRVar.

3.2.4 Variance of In-links' IDR (IDRVar). IDRVar of a website is the variance of the IDR values of the site's in-link sites.

The reason we use this feature is similar to that of IGRVar and the justification is shown in Figure 6.

$$IDRVar(a) = \sqrt{\frac{\sum_{b \in S_{in}(a,t_0)} \left(IDR(b) - IDRMean(a)\right)^2}{\left|S_{in}(a,t_0)\right|}}$$

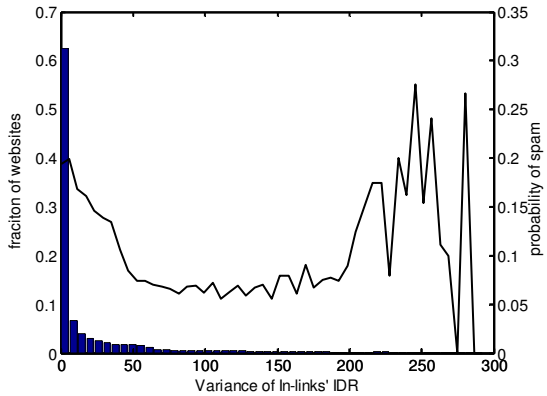

Figure 6. Probability of spam versus IDRVar.

3.3. Temporal features on the correlation of the neighbors of a website

3.3.1 Change Rate of Clustering Coefficient (CRCC). Watts and Strogatz [6] proposed clustering coefficient to measure whether or not the neighbors of a node in a graph are highly connected. Here we slightly modify their definition to fit into the current scenario. Actually, what we need consider is whether the in-links of a node are highly connected. Clustering coefficient and its change rate are defined as follows:

$$CC(a,t) = \frac{\left|\{(b,c) \in G(t) \mid b,c \in S_{in}(a,t)\}\right|}{\left|S_{in}(a,t)\right| \cdot \left(\left|S_{in}(a,t)\right| - 1\right)}$$

$$CRCC(a) = \frac{CC(a,t_1) - CC(a,t_0)}{CC(a,t_0)}$$

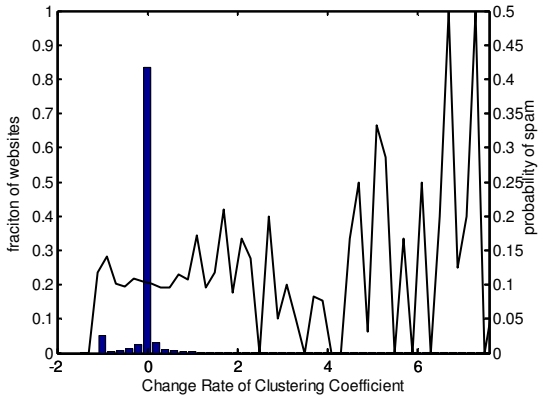

Figure 7. Probability of spam versus CRCC.

Figure 7 shows that change rate of clustering coefficient is an indicator of spam websites. As can be seen from the figure, larger change rates of clustering coefficient usually correspond to higher probabilities of being spam sites.

3.4. Other features

The above features are all related to in-link, we also define similar features on 'out-link.' For example, OGR (out-link growth rate) is a counterpart of IGR and ODR (out-link death rate) is a counterpart of IDR.

In total we have thirteen temporal features defined.

4. Experimental results

We split all the websites into a training set and a test set. Then we trained an SVM model using the tool SVM^{light} [11] with the labeled data in the training set, and then ranked the websites in the test set by the confidence scores outputted by the model. We then split the ranked websites in the test set into 'buckets' with each bucket including 100,000 websites. Each bucket has certain number of labeled data and we assume the accuracy of detecting the labeled spam websites within a bucket equals the accuracy of detecting the spam websites in the entire bucket. For comparison, we also classified the

websites in the test set using a single feature - IGR. Figure 8 shows the results for the first twenty buckets.

We can see, from Figure 8, that for the first twenty buckets the accuracy of spam detection by SVM is higher than that by IGR. Assuming that the lower bound of spam detection accuracy is 10% (i.e., 1 out of 10 sites is spam), we conclude that the performance of the SVM classifier is quite good; for most of the buckets, the accuracies are higher than 40%; some of them are even close to 60%. This result indicates that using temporal features to detect link spam is a feasible approach.

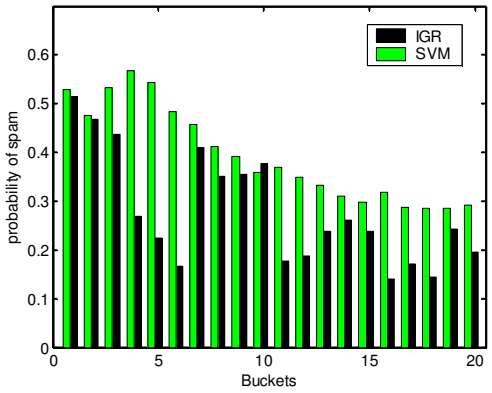

Figure 8. Spam probability of the top 20 buckets.

We further conducted qualitative evaluation on the results. We checked the top ten websites identified as spam sites by the SVM classifier. We found that among the top ten websites eight of them are true spam sites, including five link spam sites, two comment spam sites, and one site both involved in link spam and comment spam. The other two websites are suspicious ones.

We also checked the bottom ten spam websites that SVM classifier is not able to detect. We found that four of them have expired while the others have not been updated for a long time. This indicates that our method works better for detecting active spam websites, but can fail to detect those spam sites without temporal changes. To deal with the problem, we need to employ other spam detection methods.

5. Conclusions and future work

In this paper, we have proposed using temporal information in link spam detection. Specifically we have defined temporal features and have used them in an SVM model for link spam detection. The basis of our work is that link spam can be easily detected by looking at the changes of links along the time axis. Experimental results show that our claim is correct and the proposed features are indeed effective for link spam detection.

As future work, we plan to integrate our method with other existing methods to build a more powerful spam detector. We also intend to apply the same idea to *webpage graphs*, which have much larger scales.

6. Acknowledgment

We thank Wei-Ying Ma for his guidance in this project. We thank Amit Aggarwal, Rangan Majumder, and Krishna Gade for their suggestions and comments.

7. References

[1] A. A. Benczur, K. Csalogany, T. Sarlos, and M. Uher, "SpamRank - fully automatic link spam detection", In *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[2] A. L. da Costa Carvalho, P. Chirita, E. S. de Moura, P. Calado, and W. Nejdl, "Site level noise removal for search engines", In *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, Scotland, 2006.

[3] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis". In *Proceedings of the 15th International Conference on World Wide Web*, Edinburgh, Scotland, 2006.

[4] B. Wu and B. D. Davison. "Identifying link farm spam pages". In *Proceedings of the 14th International World Wide Web Conference*, Chiba, Japan, 2005.

[5] B. D. Davison, "Recognizing nepotistic links on the web", In *Proceedings of the AAAI-2000 Workshop on Artificial Intelligence for Web Search*, 2000.

[6] D. J. Watts and S. H. Strogatz, "Collective Dynamics of Small-World Networks", *Nature*, 363:202—204. 1998.

[7] D. Fetterly, M. Manasse and M. Najork, "Spam, Damn Spam, and Statistics", In *Proceedings of the 7th International Workshop on the Web and Databases*, Paris, France, 2004.

[8] D. Fetterly, M. Manasse, and M. Najork. "Detecting phrase-level duplication on the World Wide Web". In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research & Development in Information Retrieval*, Salvador, Brazil, 2005.

[9] M. Kimura, K. Saito, K. Kazama and S. Sato, "Detecting Search Engine Spam from a Trackback Network in Blogspace", Lecture Notes in Computer Science, vol. 3684, 2005.

[10] T. Joachims, L. Granka, B. Pan, and G. Gay, "Accurately Interpreting Clickthrough Data as Implicit Feedback", In *Proc. 28th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2005.

[11] T. Joachims, "Making large-Scale SVM Learning Practical". *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, 1999.

[12] Z. Gyongyi and H. Garcia-Molina, "Web spam Taxonomy", *Technical report*, Stanford Digital Library Technologies Project, 2004.

[13] Z. Gyongyi, H. Garcia-Molina and J. Pedersen, "Combating Web spam with TrustRank", In Proceedings of the 30th International Conference on Very Large Data Bases, 2004.