Study on the Structure and Behavior of Web Spam by Link Hijacking

Young joo Chung Masashi Toyoda Masaru Kitsuregawa

Institute of Industrial Science, University of Tokyo

4-6-1

Komaba Meguroku,

Tokyo, JAPAN

ABSTRACT

Today, many people try to mislead search engines to obtain a higher search ranking and bring more visitors and profits to their sites. This behavior is called web spamming. Link spamming, one of the typical methods of web spamming, manipulates link structure of the Web and deceives link-based ranking algorithms such as PageRank. Link spammers also can hijack links from good sites to their own spam sites which makes link-based ranking algorithms believe spam sites are endorsed by good sites. Although the number of link hijacked sites is small compared to that of total spam sites, hijacked links cause trust leakage so that have a significant effect on link-based ranking algorithms. In this paper, first we computed PageRank, TrustRank and Anti-trust Rank for whole web sites and selected sample hijacked sites with high trust and anti-trust scores. Then, we examined the several features, such as score distributions and types of hijacked sites.

1. INTRODUCTION

Web spamming is defined as the behavior of manipulating web page features to get a higher search ranking than that it deserves. Link spamming is one of the Web spamming techniques that manipulates the link structure of the Web to mislead link-based ranking algorithms such as PageRank [1]. For example, spammers can construct an artificially interlinked link structure, so called the spam farm, to centralize link-based importance scores. In addition to building spam farms, one of the characteristic methods of link spamming is *link hijacking*. Spammers can create links from external reputable pages to target spam pages, even if the authors of the external pages do not intend to link to them. For example, they can post comments including URLs to spam pages on public bulletin boards. These links are called hijacked links. These links do not endorse any relevance or quality of pages, so they mislead link based ranking algorithms which consider the link as human judgment about web pages.

In this paper, we examined the features of hijacked sites from a large-scale graph of the Japanese Web archive. By understanding hijacked sites and link hijacking behavior, we could understand web structures more closely and also contribute to the quality of web link-based ranking algorithms.

2. BACKGROUND

2.1. PageRank

PageRank [1] is one of the most well-known link-based ranking algorithms. The basic idea of PageRank is that a web page is important if it is linked by many other important pages. This recursive definition can be showed as following matrix equation:

$$\mathbf{p} = \alpha \cdot \mathbf{T} \cdot \mathbf{p} + (1 - \alpha) \cdot \mathbf{d}$$

where $\mathbf{p}$ is PageRank score vector, $\mathbf{T}$ is transition matrix. $T(p, q)$ is 1/the number of outlinks of q if there is a link from node q to node p, and 0 otherwise. The decay factor $\alpha < 1$ (usually 0.85) is necessary to guarantee convergence and to limit the effect of rank sink. $\mathbf{d}$ is a uniformly random distribution vector.

2.2. TrustRank and Anti-TrustRank

To demote spam pages and make PageRank resilient to link spamming, Z.Gyöngyi et al. suggested TrustRank [2]. TrustRank introduced the concept of trust for web pages. In order to evaluate trust score of the entire Web, TrustRank assigns trust score on the whole web pages and propagates scores throughout the link structure. Also, Krishnan et al. proposed Anti-TrustRank to find out spam pages [3]. As the inverse-version of TrustRank, Anti-trust Rank propagates Anti-trust score through inlinks from seed spam pages.

3. EXPERIMENT

3.1 Data Set and Seed Set

In order to analyze hijacked sites, we computed score of various versions of PageRank on a large-scale snapshot of our Japanese web archive built by a crawl conducted in May 2004. This snapshot is composed of 96 million pages and 4.5 billion links. Then, we created sites level graph of the Web where nodes represent for sites and edges represent the existence of links between pages in different sites. The site graph built from our snapshot includes 5.8 million sites and 283 million links. We call this dataset web graph in

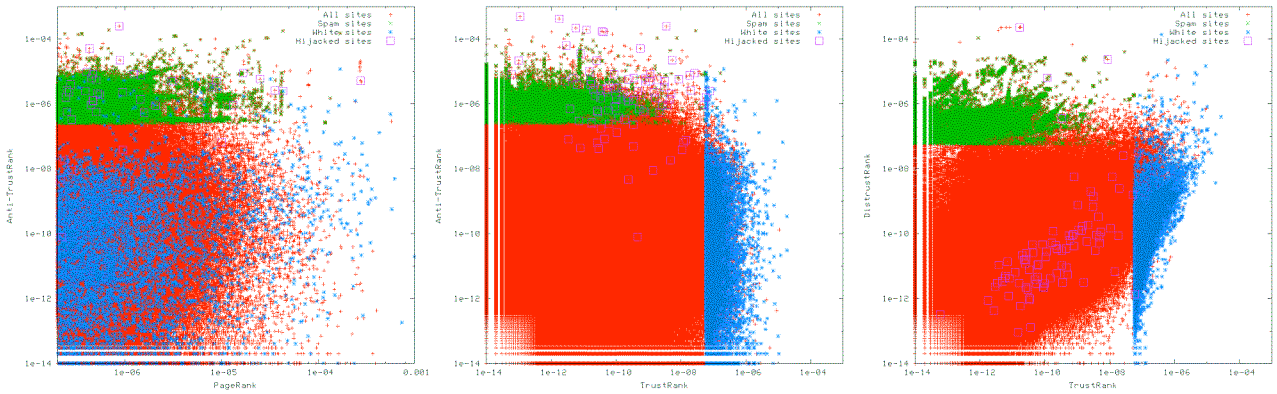

Figure 1 The score distributions of variations of PageRank. The left distribution shows the relation between PageRank and Anti-TrustRank. The middle one represents TrustRank and Anti-TrustRank. The right shows the distribution of TrustRank and DistrustRank.

this paper.

To compute TrustRank and Anti-TrustRank, we constructed trust seed set and spam seed set. We used manual and automated selection for both seed sets. As a result, we have 40,396 sites as trust sites and 580,325 sites as spam seed sites. Large spam seed set was obtained with [4].

3.2 Result

3.2.1 Types of hijacking

We obtained suspicious sites with abnormal TrustRank and Anti-TrustRank score combination. As a result, 92 hijacked sample sites with relatively high TrustRank score and high Anti-TrustRank score are founded. Then, we classified these sites into 7 types as follows.

Hijacking type	Number of sites
Blog	25
Bulletin board	20
Expired sites	20
Link register sites	8
Hosting sites	5
Server statistics	4
Normal sites having ad to spam sites	10
Total	92

Table 1 Types of Hijacking

We found out that several hijacking method exist in the real Web. In addition to posting comments on blogs or bulletin boards, spammers can buy expired domains and hijack links of normal sites pointing it. Also they can register their links on free link register sites, put links on hosting company sites as a customer and make links to spam sites by sponsoring some sites. Also, spammers can access sites with public access log statistics showing links to referrer sites frequently so that their sites are appeared in the referrer list.

3.2.2 Score distributions of PageRank variations

Figure 1 shows the distributions of several rank scores. Red, green, blue points represent all sites, spam seeds and white seeds, respectively. Purple squares are for sample hijacked sites. In first two figures, we can recognize that hijacked sites tend to be judged as spam. Their scores are similar to those of spam sites rather than white sites. Namely, hijacked sites usually get high Anti-TrustRank score. However, it seems hard to figure out a certain correlation between hijacked sites and PageRank or TrustRank scores.

Therefore, we computed a different variation of PageRank, DistrustRank. DistrustRank is similar to TrustRank, but it propagates initial scores from spam seed through outlinks. The distribution is shown in the right figure. We can see that there exists a correlation between TrustRank and DistrustRank of hijacked sites. Correlation coefficient is 0.47. When we put away 3 hijacked samples with exceptionally high DistrustScore, correlation coefficient is 0.66.

4. CONCLUSION

In this paper, we examined the types and various rank score distributions of hijacked sites. We found out there exist several hijacking methods and hijacked sites are likely to be determined as a spam site. Also, we discovered a correlation between TrustRank score and DistrustRank score that might be useful to extract hijacked sites.

5. REFERENCES

[1] L. Page, S. Brin, R. Motwani and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[2] 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*, 2004.

[3] Krishnan and R. Raj. Web Spam Detection with Anti-Trust Rank. *In Proceeding of the 2nd Adversarial Information Retrieval on the Web*, 2006.

[4] H. Saito, M. Toyoda, M. Kitsuregawa and K. Aihara. A large-scale study of link spam detection by graph algorithms In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*, 2007.