A Technique for Detecting Web Spam from a Densely Connected Directed Graph of Sites

Bingshuang Han † Masashi Toyoda ‡ and Masaru Kitsuregawa ‡

Institute of Industrial Science, the University of Tokyo

Komaba 4-6-1, Meguro-ku, Tokyo, 153–8505 Japan

E-mail: † hbshuang@tkl.iis.u-tokyo.ac.jp, ‡ {toyoda, kitsure}@tkl.iis.u-tokyo.ac.jp

ABSTRACT:

In this paper, we propose a technique for detecting link spam sites in the Web. Link spam sites attempt to deceive link-based ranking algorithms of search engines by building densely connected structure between sites.

Our method detects densely connected sets of sites from a directed graph of sites based on several patterns of directed connections, such as cycles and co-citations. We discuss which patterns are useful for detecting link spam, and show results of experiments on our Japanese web archive.

Keywords:
Densely Connected, directed graph, Link Spam, union-find

1. Introduction

In this digital information age, more and more people rely on the Internet to find all kinds of information needed. Although today's search engines can easily return millions of pages for a certain query, it is impossible for users to preview all the results. Therefore, owners of web sites expect to always be shown up on the top of result lists. This leads to the emergence of SEO (Search engine optimization) which helps web pages to acquire high ranking scores in search engines. Some examples of these techniques are: using significant titles for Web pages, giving descriptive words in the Meta tags, etc. However, there is no clear definition in the legitimation of these pages and moreover, it brings black arts such as web spamming [10] where some authors create web sites with the main purpose of misleading search engines and obtaining higher ranking than the deserved ranking .

Link spam, a newly emerging technique, takes the advantage of algorithms (i.e., HITS, PageRank), which are used to compute importance scores based on the link information.

Spammers often create link structures that help to gain undeserved high ranking scores for target pages.

Our research focuses on densely connected link structure analysis, based on one simple observation: good pages seldom points to bad ones. Therefore, the chance to detect spam pages in one densely connected cluster should be very high if we can discriminate the "good" and the "bad" ones. Our main contributions of this paper are:

1. We propose a method for detecting the web spam structure based on several patterns of connections (introduced in Section 3.1)
2. We examined appropriate connection patterns and threshold for clustering the spam sites.
3. We show the results of an extensive evaluation, based on 600 million sites and a manual examination of over 2400 sites.

The rest of this paper is organized as follows: the background and related work is introduced in Section 2. Web extraction patterns are explained in Section 3. The experimental results are shown in Section 4 and we conclude the paper in Section 5.

2. Background and Related work

2.1 Background

Usually, current search engines combine several algorithms to calculate the ranking score of pages. One of the most famous ones is PageRank which uses the link information to assign the numerical weighting to each page in the Web [4].

Another well-known algorithm for evaluation pages is HITS, which rates web pages for their authority and hub values. HITS uses two values for each page: Authority Value and Hub Value.

2.2 Related Work

While the term “spam” appeared as early as in year 1996, link spam, as one way of web spam, has acquired a highlighted position by year 2004. From its appearance, link spam attracted the attention of researchers in database, IR and Web. However, the research on the particular issues of link spam, i.e., huge amount of data, on-line processing, is still at its starting and most works focus on algorithm development to identify link spam.

Hector Geocia-Monila presented the “Link Spam Alliances” to give a detailed analysis for how spam farm can optimize web pages ranking by interconnecting each other and their results also shows the optimal structures of spam farm and quantify the potential gains in ranking score[10].

Davison showed the idea about recognizing and eliminating nepotistic links and demonstrated recognition of such links with high accuracy automatically is potential [6].

Broder and Bharat analyzed large amount of web pages and observed that the in-degree and out-degree should follow Zipfian’s law and found that “artificially” generated link farms are the outliers in the distribution [5].

Fetterly analyzed the distribution of many web page features over 429 million pages. They found the pages generated automatically are quite different from the pages authored by a human; moreover, they described several properties that help to indicate web spam pages [7, 8].

Gyongyi et al. described a new algorithm, Trustrank, to combat Web Spam [11]. The basic idea is that good pages always link to good pages and seldom point to bad ones. They first selected seed pages, and then these seed pages propagate trust weights along the hyperlinks. However, the manual selection of trusted pages creates a perceptive bias as unknown and remote websites become less visible. Wu et al proposed Topical TrustRank in 2006; they made a supplement by arranging the page the topic biased TrustRank. [14]

Wu et al. introduced one algorithm to detect link farms [13]. They first chose the seed set (spam pages) based on common link structure between incoming and outgoing links of the web pages. Then they expanded the seed set with the nodes that have too many outgoing links to the original seed set.

3. Web Spam Extraction Methods

3.1 Web model

We adapted the web model as directed graph G (V, ع) consisting of:

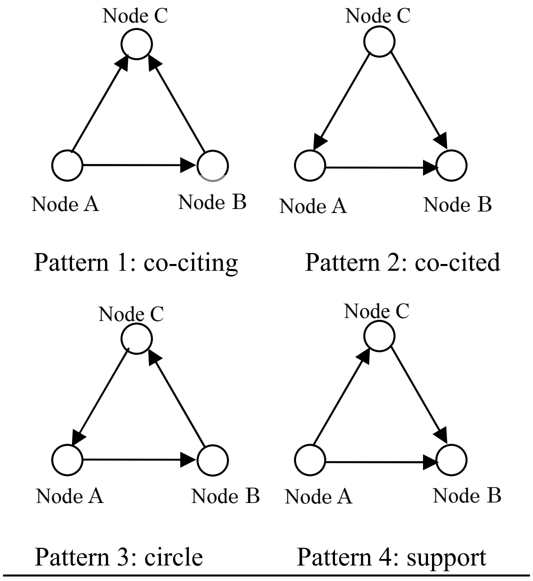
V, a set of vertices (sites)

ع, a set of edges(links) between vertices.

We use a pair of nodes (A, B) to express the link from node A to node B. For the sake of simplicity and without loss of generality, we collapse multiple links between two nodes into a single link, and also remove self links from ع.

3.2 Patterns extraction

Let us consider the link connection among three nodes, the smallest link farm alliance unit in the network. Suppose three nodes A, B, C and the links between any two nodes exist. The link direction between nodes A and B is determined, i.e., from A to B. Link directions between node C and nodes A, B are unknown. Note that there are four possible link structures for these three nodes. We give the definition of these four patterns as follows (shown in Figure 1):



Pattern 1: co-citing

Pattern 2: co-cited

Pattern 3: circle

Pattern 4: support

Figure 1: The definition of 4 patterns

Pattern 1: Co-citing. Both node A and B are co-citing to node C.
Pattern 2: Co-cited. Both node A and B are co-cited from node C.
Pattern 3: Circle. Node A, B, C has a circle link structure.
Pattern 4: Support. Node C supports the edge A to B.

The complete algorithm is shown in Figure 2. In the implementation, we extract the incoming sites and outgoing sites of both nodes A and B, and denote them as i.A (incoming sites list of A), o.A (outgoing sites list of A), i.B (incoming sites list of B) and o.B (outgoing sites list of B). To extract the four patterns, we just need to compare the corresponding sets to obtain the common shared nodes in these sets. For instance, since node C is co-cited to node A and B in Pattern 1, we should compare sets o.A and o.B and output shared node C and its number. The computation complexity of this algorithm is O (d * |E|) where d is maximum degree of nodes, and |E| is the number of edges in the Web graph.

```
Clustering method
Input:
        P          Node
        Np.in      The in-link number of P
        Np.out     The out-link number of P
        E (A, B)   The edge from A to B

●   Cluster node C (E.g Pattern 1)
    For each edge in G (P, E)
      edge =(A->B) ( $A, B \in P$ )
      o.B   =out-links (B)
      o.A   =out-links (A)
      num =number of nodes shared between
            ( i.B, i.A)
Output:
      (A, B, num)
```

Figure 2: The algorithm for calculating the degree of connections between A and B.
(In the case of Pattern 1)

3.3 clustering based on union-find method

In the previous step, we obtained the results of node A and B's common sharing nodes in each pattern. Moreover, we had the distribution of the shared nodes for each pair of nodes in 4 patterns. As we focus on the densely connected directed web graph, we made a setting of threshold for merging pairs of nodes based on union-find algorithm. For instance, in co-citing pattern, a pair of nodes<A, B> shares N1 nodes, and a pair of nodes <A, D>shares N2 nodes. If both N1 and N2 are bigger than the threshold N, then we merge <A, B> and <A, D> into one cluster (A, B, D). Figure 3 shows an example of edge <A, B> shares 4 nodes in pattern Co-citing.

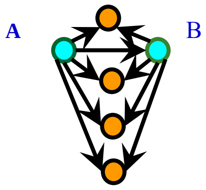


Figure 3: an example of <A, B> shares 4 nodes

Shared Nodes Size	Pair of node <A, B> included(edges)			
	Co-citing	Co-cited	Circle	Support
10	95.85%	95.23%	84.45%	97.39%
50	77.78%	79.58%	74.46%	81.72%
100	71.33%	73.70%	69.65%	73.96%
500	11.84%	12.79%	9.86%	11.60%

Table 1: Edges included in the results of pattern extraction

Shared Nodes Size	Sites supposed to be spam			
	Co-citing	Co-cited	Circle	Support
10	18.45%	19.20%	11.15%	19.63%
50	8.48%	9.00%	6.64%	10.46%
100	5.03%	5.97%	4.38%	6.58%
500	1.17%	2.17%	1.93%	2.14%

Table 2: Nodes included in the results of pattern extraction

3.4 Number of false positives-sample examination

For the sake of validate the quality the result in dis-including non-spam sites; we did false positives-sample examination.

First, we manually collected some non-spam sites from the sites list in original dataset, in the order of descending out-degree, and made them into a white-site list. Second, compare the sites supposed to be spam sites in the result and the sites in white-site list to identify how many labeled white-sites are included in the result.

By reviewing the results, certainly, the less number of labeled white-sites included, the better quality of result. We can determine the appropriate threshold *N* and the appropriate pattern for spam detection.

4. Experiments

4.1 Data set

The data used in our experiment is a large-scale crawling of Japanese web sites collected in May 2004, including 5.8 million of sites and 283 million of edges. The format of nodes (sites) has three tiers (i.e. http://A/B/C).

In order to see the how the precision in distinguishing spam sites changes when filter the low-degree sites, we made the second dataset which is consist of densely connected sites only, by filtering the sites of more than 100 in-links or 100 out-links.

4.2 Results for extraction based on 4 patterns

According to the patterns for extraction which were introduced in section 3.2, we counted the number of the shared nodes for each pair of nodes. Figure 4 illustrates the result of this step in experiment - the distribution of connections degree between each pair <A, B>.

We can see some sites, which are densely connected each other with the size of shared nodes more than 1000, can also be extracted.

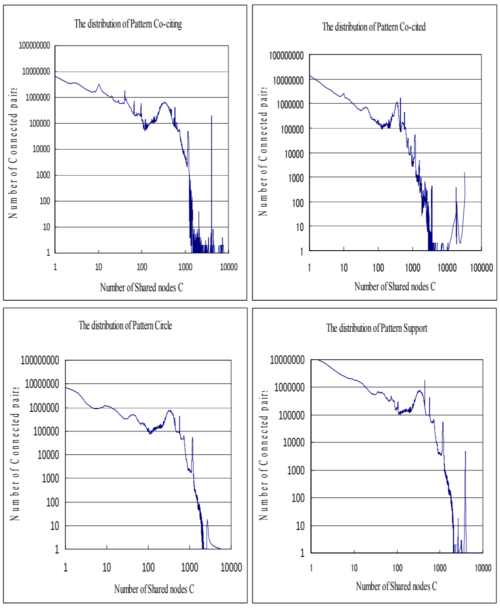


Figure 4: The distribution of degree of connections between A and B in 4 patterns

4.3 Results for Union-find cluster

Based on the previous step, we set the thresholds of shared nodes size for merging nodes A and B as 10, 50, 100, and 500 in union-find based clustering, to see how cluster size distribution changes with different threshold. Figure 5 shows the experimental result of the cluster size distribution based on union-find algorithms with different thresholds. From the figure, we can see some big clusters can be extracted. E.g. the distribution of Co-citing pattern with threshold 10 has 6898 clusters within which the biggest cluster contains 23634 sites. This shows our patterns not only can extract the completed link structure like cliques but also can extract some relatively dense sub-graphs in massive graphs.

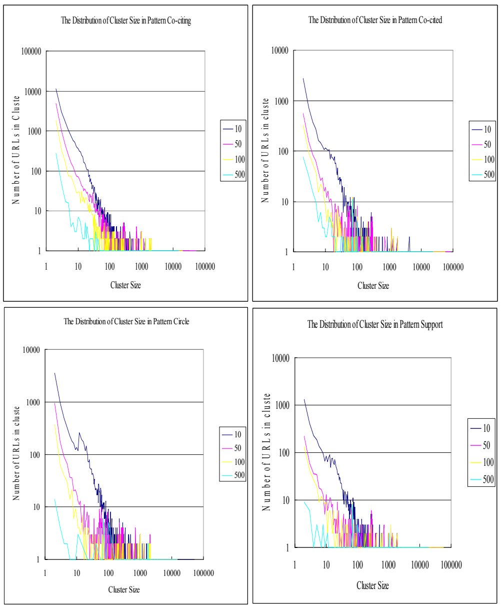


Figure 5: The cluster size distribution based on union-find algorithm with different thresholds

Table 1 and 2 shows the linkage information covered in the pattern extraction and the sites which are supposed to be spam. It is easily to say that small amount of sites have a majority of links (E.g. in Co-citing pattern with threshold 100, 5% of sites hold the 70% linkage information in the whole web graph).

4.4 Number of false positives-sample examination

We manually selected 153 white (non-spam) sites from the top linked sites, and counted how many labeled white sites would be included in each pattern with regard to different thresholds. Table 3 shows this result. We can observe that the pattern Co-citing and the pattern Circle have better performance in dis-including labeled white sites.

Considering the balance between the precision and coverage rate in spam detection, therefore, 100 is a suitable threshold. In other words, when the threshold becomes bigger, the precision becomes higher, but the coverage in spam sites becomes lower.

Pattern N	10	50	100	500
Co-citing	64	28	19	5
Co-cited	152	152	149	120
Circle	72	34	22	5
Support	150	152	151	87

Table 3 the labeled white sites included in 4 patterns with different threshold

4.5 Results for spam detection

We intended to check the precision of spam sites detection to confirm which pattern is the best one for spam detection. The implementation detail is described as follows: For each pattern, we randomly chose 100 clusters, and manually inspected the content of one site in each cluster. To eliminate the effects caused by the non-uniform distribution of the spam cluster size, we just kept the results of clusters whose sizes are more than 10 and made them into figures. In order to display the result distinctly, we defined the classification of the subjects in:

- Non-Spam: Sites offer useful information, including personal sites, corporation sites and sites of public institutions.
- Link Directory & Sales Promotion: the content of sites just consists of link information and the advertisement about product, such as discount information and etc.
- Pornographic sites
- Unsure: unknown language

Figure 6, 7 and 8 show the results of manually spam detection. In figure 5, we focus on pattern Co-citing and Circle with thresholds in 10, 50, and 100. Obviously, depending on the increasing thresholds, the precision of spam extraction become higher.

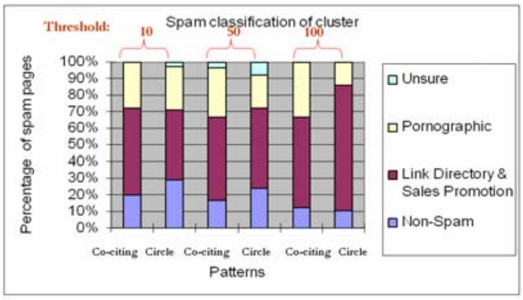


Figure 6: Spam classification with different thresholds in Co-citing and Circle pattern (Dataset 1)

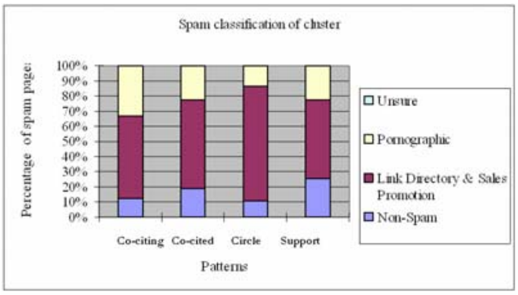


Figure 7: Spam classification of clustering with threshold 100 in each pattern (Dataset 1)

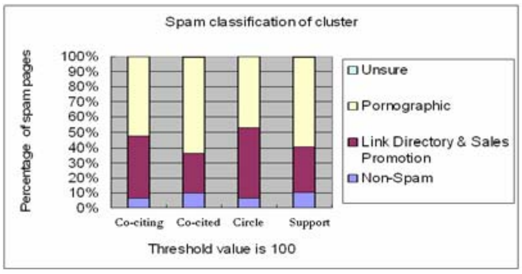


Figure 8: Spam classification of clustering with threshold 100 in each pattern (Dataset 2)

Figure 7, 8 are the comparison of the spam detection in two datasets. We set the threshold to 100, and check the precision in 4 patterns. We found filtering the low-degree sites can provide higher precision in spam extraction.

Figure 8 shows the best performance of all the results, in Co-citing pattern with threshold 100(with dataset 2). More than 95% of the sites extracted are real spam.

4.6 Analysis of results

We can see from Table 3 and Figure 8 that Pattern Co-citing and Circle provide better performance as the numbers of non-spam sites are comparatively smaller.

A theoretical explanation of the reason is: spam sites could have out-links pointing to non-spam sites to boost hub values as shown in Figure 9. In the upper case, node A and node B are spam sites, and they share many nodes in Co-citing pattern. We do clustering of node A and B based on union-find algorithm. Therefore, the precision of spam extraction in Co-citing pattern provides better performance. In the latter case, node A and B are non-spam sites, and they are linked by many spam sites, we can say, node A and B share many nodes in Co-cited pattern. Therefore, the clustering result of node A, B is worse in spam extraction.

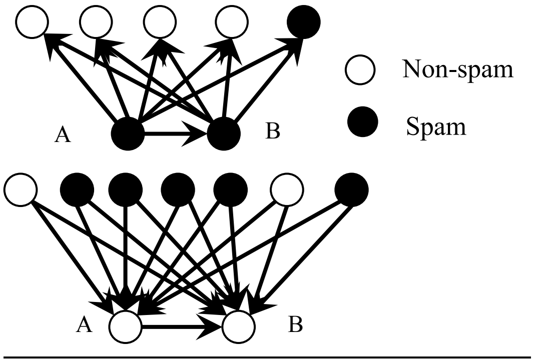


Figure 9: Cases of spam sites point to non-spam site

5. Conclusions

This paper presented a technique to detect web spam from a densely connected directed graph of sites. By applying union-find algorithm and clustering based on 4 basic patterns, we are able to identify link spam efficiently. Our experimental results demonstrated that we can identify most of link spam. Furthermore, pattern co-citing and pattern circle have better performance to avoid mistaking non-spam sites for spam sites.

References

[1] http://www.graphviz.org/ Information:

[2] Andras A. Benczur, K. Csalogany, T. Sartos and M. Uher. "SpamRank-Fully Automatic Link Spam Detection". In 1st International Workshop on Adversarial Information Retrieval on the Web, 2005.

[3] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, R. Baeza-Yates. "Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection". The future of Web Search *Workshop*, May 19-20, Barcelona, Spain, 2006

[4] M. Bianchini, M. Gori and F. Scarselli "Inside Pagerank" ACM Transactions on Internet Technology (TOIT) Volume 5 , pages 92-128, USA ,Feb 2005,

[5] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval, pages 104-111, Melbourne, AU, 1998

[6] Brian D. Davison. "Recognizing nepotistic links on the web". In AAAI-2000 Workshop on Artificial Intelligence for Web search, pages 23-28, Austin, USA, July 30, 2000

[7] D. Fetterly, M. Manasse and M. Najork and J. Wiener. "A large-scale study of the evolution of web pages". In Proceedings of the 12th International World Wide Web conference (WWW), Budapest, Hungary, 2003

[8] D. Fetterly, M. Manasse and M. Najork. "Spam, damn spam and statistics-Using statistics to locate spam web pages". In Proceedings of the 7th International Workshop on the Web and database (WEBDB), Paris, France, 2004

[9] Z. Gyongyi and H. Garcia-Monlina. "Web Spam taxonomy". In Proceedings of the 1st International Workshop on Information Retrieval on the Web (AIRWEB), 2005

[10] Z.Gyongyi and H. Garcia-Molina. "Link Spam Alliances". In Proceedings of International Conference on Very large Database (VLDB), Trondheim, Norway, 2005.

[11]Z. Gyongyi, H. Garcia-Monlina and J. Pedersen. "Combating web spam with TrustRank". In Proceedings of International Conference on Very large Database (VLDB), Toronto, Canada, 2004.

[12] H. Ono, M. Toyoda and M. Kitsuregawa "Identifying Web Spam by Densely Connected Sites and its Statistics in a Japanese Web Snapshot" DESW 2005, Japan.

[13] B. Wu and B. D. Davison. "Identifying Link Farm Spam Pages". WWW2005, Chiba, Japan, May 10-14, 2005

[14] B. Wu, V. Goel and B. D. Davison. "Topical TrustRank: Using Topicality to Combat Web Spam". WWW2006, May 23-26, Edinburgh, Scotland, 2006