Analysis of Web Spam Structure Using Recursive Strongly Connected Component Decomposition

Young-joo Chung Masashi Toyoda Masaru Kitsuregawa

Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, Japan

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

Abstract Addressing Web spam is a critical issue for today's search engines. In this paper, we studied a structure of the Web spam using recursive strongly connected component (SCC) decomposition. Spam pages are likely to construct densely connected structures; consequently, SCC decomposition would catch the spam structure of the Web efficiently. Also, by recursive SCC decomposition algorithm with node filtering, denser spam structures would be extracted. We applied SCC decomposition algorithm to our Japanese Web archive crawled in 2004, and separate many large components from a core, the largest component. After this, SCC decomposition algorithm performed again to nodes in the core which have degrees over a threshold. We did this decomposition recursively with increasing threshold. As a result, we found out large link farms in each iteration and this trend continues until at least 10 iterations. We investigated large links farms whether they are spam or not by their URL characteristics. The result showed almost large link farms are spam.

Keyword Web spam, Link analysis, Strongly connected components

1. INTRODUCTION

Web spamming is the attempt to boost a search ranking of a target page not by improving the quality of a page but by manipulating the features of a page. Since many people rely on search engines to buy from daily goods to a house, site owners are making a serious effort to attract traffic which is connected directly to revenue. A study of 2006 reported that 13.8% of English Web pages were spam [1], and in many cases, spam pages are successful way to boost site rank. [2]

Repeating popular keywords which is not related with page contents (*term spamming*), or generating numerous links pointing to a target pages (*link spamming*) are typical techniques to manipulate a ranking. Particularly, link spammers create a densely connected link structure, a *link farm*, to mislead search engines. Although many efforts to detect and demote spam have been made for a long time, Web spam still exists and spamming techniques evolve as the contents of Web grow and diversify.

In this paper, we study an overall spam structure in a large host graph of the Japanese Web crawled in 2004. By understanding the spam structure, we could invent more efficient anti-spam strategies. In our previous work [6], we showed that most of large SCCs (except for the largest one, so called the core) are link farms. In this paper, we proposed a different approach for finding link farms in the core. We prune small degree nodes from the core, and recursively apply SCC decomposition to the pruned core in order to extract link farms from the host graph.

The rest of this paper proceeds as follows. In Section 2, we review some previous works related with our study. Section 3 describes our data set. In Section 4, the experimental result is presented. Finally, we summarize and conclude our works in Section 5.

2. PREVIOUS WORK

Link-based ranking algorithm such as PageRank [4] and HITS [3] are main targets of link spammer. Since these algorithms consider a link to pages as an agreement to that page, spammers create numerous false links and construct an artificially interlinked link structure, so called a spam farm, to centralize link-based importance to their own spam pages [10].

Several approaches have been suggested for the purpose of detecting and demoting link spam. To understand the characteristics of Web spamming, Gyöngyi et al. described various Web spamming techniques in [9]. Optimal link structures to boost PageRank scores are also studied to grasp the behavior of Web spammers [10]. Fetterly et al. found out that outliers in statistical distributes are very likely to be spam by analyzing statistical properties of linkage, URL, host resolutions and contents of pages [7]. To demote link spam, TrustRank [11] is introduced which is a biased PageRank where rank score start to propagate from a seed set of good pages through outgoing links. By this, we can lower rank scores of spam pages. Optimizing the link structure is another approach to demote link spam. Carvalho et al. proposed the idea of noisy links, a link structure that has a negative impact on the link-based ranking algorithms [13]. Qi et al. also estimated the quality of links by similarity of two pages [14]. To detect link spam, Benczur et al. introduced SpamRank [10].

SpamRank checks PageRank score distributions of all in-neighbors of a target page. If this distribution is abnormal, SpamRank regards a target page as a spam and penalizes it.

Saito et al. employed Graph algorithm [6] to detect link spam. They decomposed the Web graph into strongly connected component and discovered large components are very likely to be spam. Link farms in the core were extracted with minimum flow cut which use spam seed obtained by maximal clique enumeration. This work is similar to ours in the respect that both apply SCC decomposition on the Web, but we introduced recursive SCC decomposition to extract spam structure in the core instead of minimum cut.

3. DATSET

We performed experiments on a large-scale snapshot of our Japanese Web archive built by a crawling conducted in May 2004. Our crawler is based on breadth-first crawling [15], except that it focuses on pages written in Japanese. We collected pages outside the .jp domain if they were written in Japanese. We used a Web site as a unit when filtering non-Japanese pages. If it could not find any Japanese pages on the site within the first few pages, the crawler stopped collecting pages from a site. Hence, this dataset contains fairly amount of English or other language pages. The amount of Japanese pages is estimated to be 60%. 96 million pages and 4.5 billion links are included in this snapshot. Our crawler does not have explicit spam filter while it detects mirror servers and tries to crawl only representative servers. Therefore, our archive includes spam sites without mirroring.

In this paper, we will use a host graph, where each node is a host and each edge between nodes is a hyperlink between pages in different hosts. The properties of our Web snapshot are shown in Table 1.

Table 1 The Properties of the host graph

# of nodes(host)	2,978,223
# of edges	67,956,304

4. EXPERIMENTS

4.1. Strongly Connected Component Decomposition with Degree Filtering

In order to extract link spam efficiently, we decomposed the host graph into strongly connected components (SCCs), where every pair of nodes has a directed path between them. SCCs of a directed graph are maximal strongly connected subgraphs. It is known that the SCC decomposition of the whole Web graph produces the largest SCC (so called core) with about 30% of all nodes, and many smaller SCCs. Since spam pages construct a densely connected link structure [8], and links between spam and normal pages seldom exist, it can be expected that spam pages forms a large link farm. Our previous work [6] confirmed that the 95% of large SCCs around the core are spam link farms, but we could not efficiently find denser link farms left in the core. In this paper, we introduce a recursive decomposition of the core. After decompose the whole Web, we filter out nodes whose indegree and outdegree are smaller than 2, then apply SCC decomposition algorithm to left nodes in the core. In the third decomposition, we increase the degree threshold to 3, then apply SCC decomposition to the core obtained by the second decomposition. This process is repeated recursively while we can obtain large SCCs in the results. Here is terminology we will use in this paper.

Core the largest strongly connected component
Level n node a node in level *n* - 1 core and with both in and outdegrees over *n*.
Level n core the largest component, or core, obtained after SCC decomposition of level *n* nodes.

A detailed result for the decomposition of different levels is shown in Table 2. The percentage of a core size increases drastically between level 1 and level 2. This implies that in the core of the Web, the connectivity of nodes is strong and hard to break.

Table 2 The result of SCC decomposition

Level	1	2	5
# of node	2,978,223	556,190	302,613
# of SCC	1,888,550	9,055	612
Size of the core (%)	749,166 25.15	520,554 93.60	301,120 99.51

Figure 1 illustrates the connectivity of components in the first and second level. The left figure shows the result of level 1 decomposition and the right one is that of level 2 decomposition. A big gray node, a black node and a white node represent a core, a SCC with over 100 nodes and a smaller SCC that connects large components, respectively. The size of node describes the number of hosts included in the SCC. Two SCCs are connected by a directed edge when hyperlinks exist between hosts in SCCs at both ends. Each edge starts from the thick end and

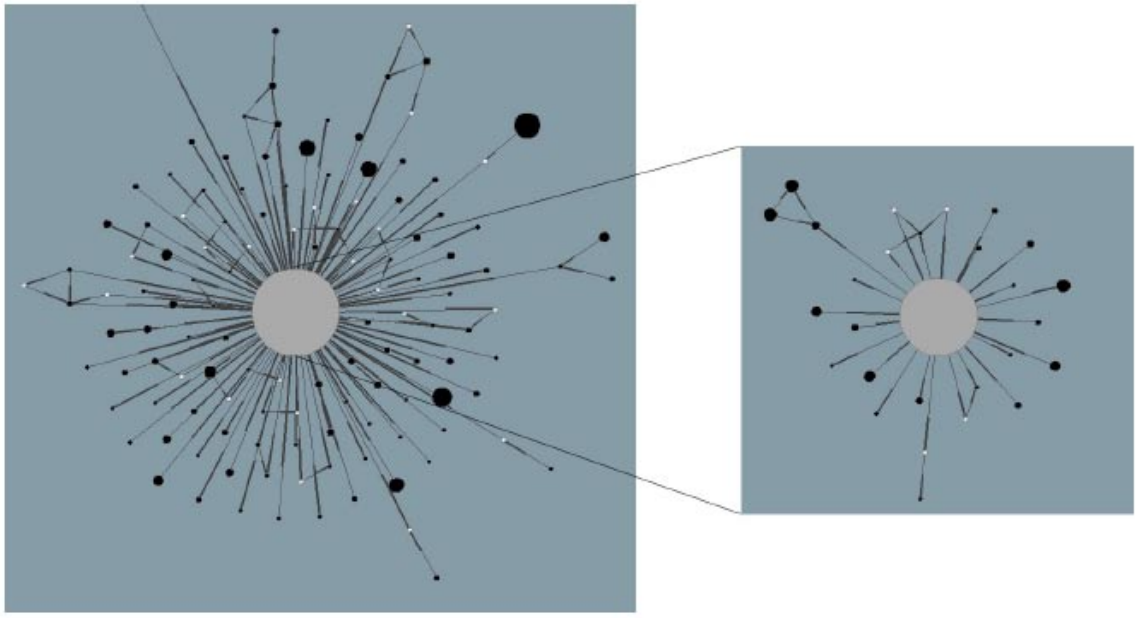

Figure 1 Component connectivity of the entire nodes and the level 2 nodes

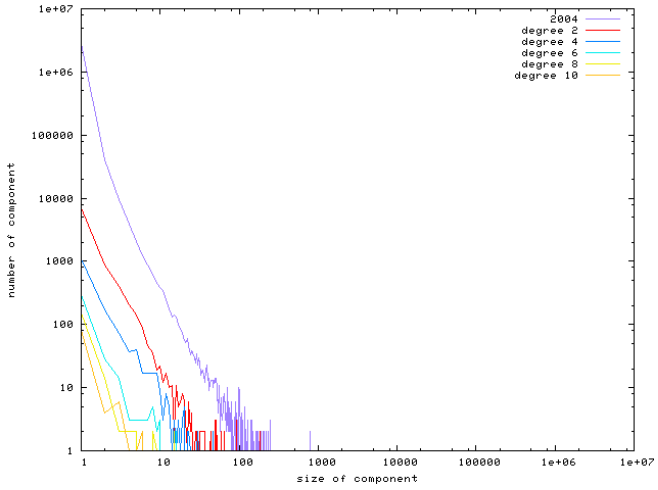

Figure 2 Component size distribution of 2004

goes to the thin end.

When comparing left and right figures, we can see the similar structure appears in the decomposition result of both the entire hosts and level 2 hosts. In addition to, most large components are directly connected to the core. We checked how the level 1 components are connected to the level 2 components. For total 49 components that include over 100 hosts, 17 components are linked by non-spam host in the core, via link hijacking. The details are demonstrated in Table 3. Links from a normal host to a spam host are hijacked links. Unknown hosts are written in unrecognizable languages like German and Spanish. Note that there is one link from a normal host to a normal one. We found out that the source host is a Japanese host while the destination host is Korean one. Korean hosts constructed a large SCC outside of the core.

Table 3 Type of links between the core and large SCC

Source	Destination	# of Large SCC around core
Normal	Spam	17
Spam	Spam	25
Normal	Normal	1
Unknown	Unknown	6
Total		49

4.2. Size Distribution of Strongly Connected Components

The component size distribution of nodes in different levels is illustrated in Figure 2. As Figure indicates, the size distribution of components obeys the power law, which agrees with the observation in Broder et al [8]. Moreover, we can observe that distributions of SCCs obtained by the decomposition of nodes in different levels also show similar distribution, which suggests the self-similar structure of the host graph. Moreover, an abnormal distribution appears at the tail of each graph. Such phenomenon is clearer in large components with over 100 hosts. We measured their spamicity and discovered these components are spam with high probability. The detail of spamicity measurement will be explained in Section 4.3.

4.3. Spamicity of Strongly Connected Components

As spamicity measurement, we used hostname properties based on the study of Fetterly et al [7]. We used two metrics; hostname length and spam words in a hostname. Average hostname length of members and the percentage of member with a hostname containing spam words were computed. Spam words were obtained by following step; we extracted hostnames from SCCs of which size is over 1000. These hostnames are split into words by non-alphabetic characters, such as periods, dashes and digits. We made a frequency list of these words and manually chose 114 spam words with high frequency from it. This spam word list contains English, Spanish, Italian, French and Japanese spam words so that it could cover most spam hostnames in various languages. We judge a hostname spam if it contains at least one spam word. If the first field of hostnames contains only non-alphabetic words such as dashes and digits, those hostnames are also regarded as spam. Then, the ratio of spam members of a component was obtained, by dividing the number of spam hostnames with the total number of hostname in a component. For all nodes, the average hostname length was 24.25 characters, and the percentage of hostnames that contain spam words are 8.97%.

The results of measurement for hosts of different levels are demonstrated in Figure 3 and 4. Log-scale is used on x axis for the size of component. In the each level decomposition, spamicity of SCCs except the largest component was examined. We can observe that as the size of components increases, the hostname length and spam word ratio also increase in level 1 and level 2. Note that the spamicity of large component in level 4 is very low, so we investigated manually and found out hostnames in these components are also spam, which are very short and consist of a series of spam words without any non-alphabetic characters. (e.g. "www.jesslysex.com") With this result, it could be said that most components with relatively large size (especially, over 100) have very high spamicity. This corresponds to the result of [6]. As for SCCs from deep level decompositions, although overall spamicity decreased, large components still have high spamicity. Since some large components with low spamicity appeared, we assessed them manually and found out all of them are actually spam. We discovered hostnames in those components are very short and consist of a series of spam words without any non-alphabetic characters. Table 4 shows the number of component with

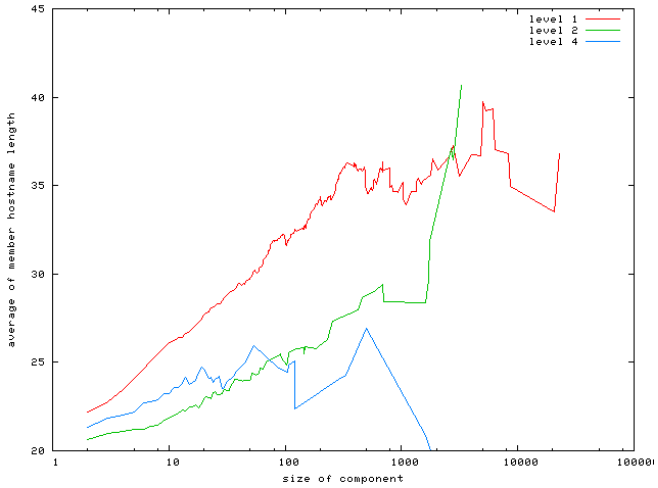

Figure 3 Average member hostname length of components in different levels

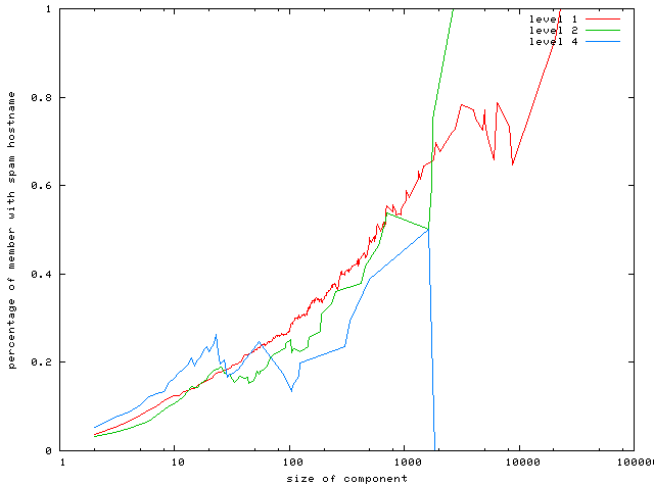

Figure 4 Rate of members with a spam hostname of components in different levels

over 100 hosts and the number of hosts in those components from level 1 to level 5. The percentage of total hosts included in large components to entire hosts is 7.2%.

To confirm that whether the tendency that large component is likely to be spam continues to the deep level of the core, we investigated components with size over 100 in SCC decompositions results of from level 5 to level 10. Table 5 indicates the result. We can see that such a trend remains even when we perform SCC decomposition on nodes of deeper level.

Table 4 Number of large components and hosts in them

Level	1	2	3	4	5
#component	228	24	7	9	2
# of host	182285	18650	9306	5032	242

Table 5 Number of spam components among large components with size over 100, in deep level

Level	5	6	7	8	9	10
Spam / Total	2/2	1/2	1/2	1/1	2/2	0/0

5. SUMMARY

In this paper, we studied the overall link-based spam structure in a large scale Web archive. These results could be useful for removing major link farms and improve the quality of a Web search result. We proposed recursive SCC decomposition with node filtering as a method for extracting denser link farms in the deeper of the core. We showed that in each iteration, almost all large components that contain more than 100 nodes turned out to be a spam farm. Using this method, we could extract about 7.2% of all hosts as link farms.

REFERENCES

[1] Alexandros Ntoulas, Marc Najork, Mark Manasse and D. Fetterly. “Detecting Spam Web Pages through Content Analysis”, the 15th international conference on World Wide Web. 2006.

[2] C. Castillo, D. Donato, A. Gionis, V. Murdock and F. Silvestri. “Know your neighbors: Web spam detection using the Web topology”, Proc. the 7th SIGIR, 2007.

[3] J. M. Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, Proc. the ACM-SIAM Symposium on Discrete Algorithms, pp.668-677, 1998.

[4] S. Brin and L. Page, “The anatomy of a large-scale hypertextual Web search engine”, Proc. the 7th international conference on World Wide Web, 1998.

[5] R. Kumar, P. Raghavan S. Rajagopalan and A. Tomkins. “Trawling the Web for Emerging Cyber-Communities”, Proc. the 8th international conference on World Wide Web, 1999.

[6] H. Saito, M. Toyoda, M. Kitsuregawa and K. Aihara. “A large-scale study of link spam detection by graph algorithms”, Proc. the 3rd international workshop on Adversarial information retrieval on the Web, 2007.

[7] D. Fetterly, M. Manasse and M. Najork. “Spam, damn spam, and statistics: using statistical analysis to locate spam Web pages”, Proc. the 7th International Workshop on the Web and Databases, 2004.

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan. S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. “Graph structure in the Web”, Computer Networks, 2000, pp. 309-320.

[9] Z. Gyöngyi and H. Garcia-Molina. “Web spam taxonomy”, Proc. the 1st international workshop on Adversarial information retrieval on the Web, 2005.

[10] Z. Gyöngyi and H. Molina. “Link Spam Alliance”, Proc. the 31st international conference on Very large Data Bases, 2005.

[11] Z. Gyöngyi, H. Garcia-Molina and J. Pedersen. “Combating Web spam with TrustRank”, Proc. the 30th international conference on Very Large Data Bases, 2004.

[12] A. Benczur, K. Csaloganym T. Sarlos, M. Uher. “SpamRank-fully automatic link spam detection”, Proc. the 1st international workshop on Adversarial information retrieval on the Web, 2005.

[13] A. Carvalho, P. Chirita, E. Moura and P. Calado. “Site level noise removal for search engines”, Proc. the 15th international conference on World Wide Web. 2006.

[14] X. Qi, L. Nie and B. D. Davison. “Measuring similarity to detect qualified links”, Proc. the 3rd international workshop on Adversarial information retrieval on the Web, 2007.

[15] M. Najork and J. L. Wiener. “Breadth-first crawling yields high-quality pages”, Proc. the 10th international conference on World Wide Web, 2001.