IPSJ Transactions on Databases Vol. 4 No. 2 158–171 (July 2011)

Regular Paper

A Study of Link Farm Evolution Using a Time-series of Web Snapshots

Young-joo Chung,[†1,*1] Masashi Toyoda[†1]
and Masaru Kitsuregawa[†1]

Web spamming has emerged to deceive search engines and obtain a higher ranking in search result lists which brings more traffic and profits to web sites. *Link farm* is one of the major spamming techniques, which creates a large set of densely inter-linked spam pages to deceive link-based ranking algorithms that regard incoming links to a page as endorsements to it. Those link farms need to be eliminated when we are searching, analyzing and mining the Web, but they are also interesting social activities in the cyberspace. Our purpose is to understand dynamics of link farms, such as, how much they are growing or shrinking, and how their topics change over time. Such information is helpful in developing new spam detection techniques and tracking spam sites for observing their topics. Especially, we are interested in where we can find emerging spam sites that is useful for updating spam classifiers. In this paper, we study overall size/topic distribution and evolution of link farms in large-scale Japanese web archives for three years containing four million hosts and 83 million links. As far as we know, the overall characteristics of link farms in a time-series of web snapshots of this scale have never been explored. We propose a method for extracting link farms and investigate their size distribution and topics. We observe the evolution of link farms from the perspective of size growth and change in topic distribution. We recursively decomposed host graphs into link farms and found that from 4% to 7% of hosts were members of link farms. This implies we can remove quite a number of spam hosts without contents analysis. We also found the two dominant topics, “Adult” and “Travel”, accounted for over 60% of spam hosts in link farms. The size evolution of link farms showed that many link farms maintained for years, but most of them did not grow. The distribution of topics in link farms was not significantly changed, but hosts and keywords related to each topic dynamically changed. These results suggest that we can observe topic changes in each link farm, but we cannot efficiently find emerging spam sites by monitoring link farms. This implies that to detect newly created spam sites, monitoring current link farm is not enough. Detecting sites that generate links to spam sites would be an effective approach.

1. Introduction

The Web has become a major source of information and a place for commercial activities for the last two decades. Many people now access the Web via search engines such as Google, Yahoo! And MSN to get knowledge, reserve hotels, and buy daily product. While there are over one trillion URLs on the Web [1)], the half of users look at no more than the top five sites in search result lists [2)]. In this situation, it is essential to obtain a high ranking in search results, which leads to increase in visitors and profits. As a result, some people started manipulating pages' contents and link structures to mislead search engines and boost their rankings. This behavior is called *web spamming*, and manipulated pages are called *spam pages*.

Spammers use various techniques for manipulating textual contents and the link structure. They insert popular keywords into their pages and copy relative documents from other sites to make their sites look useful. They also create a *link farm*, a densely connected structure which consists of many inter-linked spam pages, to increase the number of incoming links and deceive link-based ranking algorithms such as PageRank [3)] which regard incoming links as endorsements to that pages.

Addressing web spam is critical not only for search engines but also for web analysis applications based on web archives, since spamming techniques confuse various web analysis. For example, when we use link-based community extraction methods such as HITS [4)] and trawling [5)], results would include many link farms. Artificially stuffed popular keywords can contaminate the result of time-frequency analysis of terms in the Web. Addressing web spam is also challenging because new spam pages are being continuously created to avoid new anti-spamming techniques and to advertise new products. For example, spammers started inserting short text segments copied from various sites to avoid document copy detection techniques. They also continue creating massive pages advertising new drugs and products, which have not yet known to spam filters.

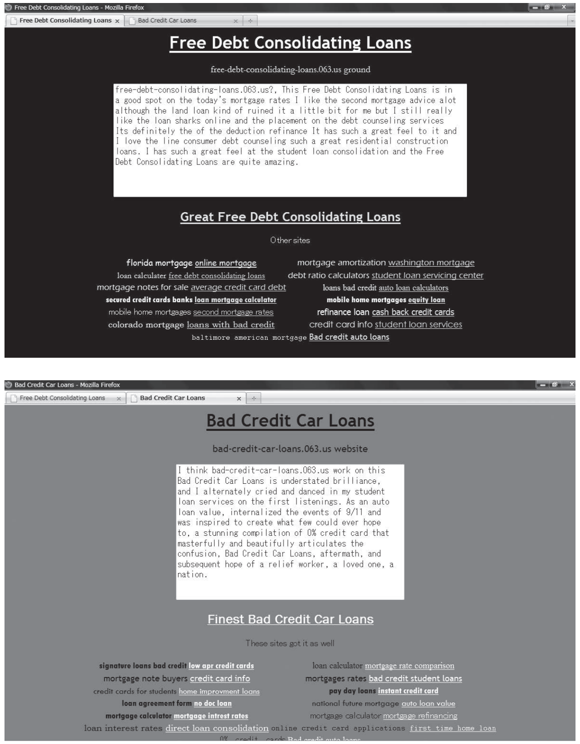
Spam pages need to be eliminated when we are searching, analyzing and mining the Web, but they are also interesting social activities in the cyberspace. In this paper, we focus on link farm and study dynamics of link farms, such as, how

[†1] Institute of Industrial Science, the University of Tokyo
[*1] Presently with Rakuten Institute of Technology, Rakuten Inc.

much they are growing or shrinking, and how their topics change over time. Such information is helpful in developing new spam detection techniques and tracking spam sites for observing their topics. Especially, we are interested in where we can find emerging spam sites that is useful for updating spam classifiers. We use large-scale Japanese web archives for three years containing four million hosts and 83 million links. As far as we know, link farms in a time-series of snapshots of this scale have never been explored.

In previous work [6], we extracted link farms in a single Web snapshot by applying a strongly connected components (SCC) decomposition algorithm to the Web graph. We showed that except for the largest SCC (i.e., the core), almost all large SCCs were link farms. We, however, could not efficiently extract link farms in the core. In this paper, we improve the SCC decomposition algorithm to extract more densely-connected link farms in the core. That is, we prune nodes with small degrees from the core and recursively apply the SCC decomposition algorithm to the pruned core with increasing a degree threshold. We extracted large SCCs for at least 10 iterations and showed that these SCCs in the core were also likely to be link farms. We found that from 4% to 7% of all hosts were members of link farms. This implies that we can remove quite a number of spam hosts from web archives only based on the link structure.

After extracting link farms, we classify topics of spam hosts in them based on URLs. We observed that spammers construct a link farm using spam hosts having URLs and contents related to the same topic. For example, **Fig. 1** shows two spam pages in one link farm. URLs of these hosts are `"free-debt-consolidating-loans.063.us"` and `"bad-credit-car-loans.063.us"`, respectively. Both hosts contain many similar keywords such as `loan`, `credit` and `debt`, which implies their topic is “Finance”. Based on this observation, we assume that a relatively small link farm consists of pages about the same topic. We select seven spam topics that are heavily targeted by spammers based on a manual investigation of small link farms and categorize spam hosts in link farms into these seven topics using URLs. We showed that adding URLs from link farms can improve the classification result. We found that two dominant topics, “Adult” and “Travel”, account for over 60% of spam hosts in link farms.


Free Debt Consolidating Loans - Mozilla Firefox
Free Debt Consolidating Loans
Bad Credit Car Loans

Free Debt Consolidating Loans

free-debt-consolidating-loans.063.us ground

free-debt-consolidating-loans.063.us?, This Free Debt Consolidating Loans is in a good spot on the today's mortgage rates I like the second mortgage advice alot although the land loan kind of ruined it a little bit for me but I still really like the loan sharks online and the placement on the debt counseling services Its definitely the of the deduction refinance It has such a great feel to it and I love the line consumer debt counseling such a great residential construction loans. I has such a great feel at the student loan consolidation and the Free Debt Consolidating Loans are quite amazing.

Great Free Debt Consolidating Loans

Other sites

florida mortgage online mortgage   mortgage amortization washington mortgage
loan calculater free debt consolidating loans   debt ratio calculators student loan servicing center
mortgage notes for sale average credit card debt   loans bad credit auto loan calculators
secured credit cards banks loan mortgage calculator   mobile home mortgages equity loan
mobile home mortgages second mortgage rates   refinance loan cash back credit cards
colorado mortgage loans with bad credit   credit card info student loan services
baltimore american mortgage Bad credit auto loans

Bad Credit Car Loans - Mozilla Firefox
Free Debt Consolidating Loans
Bad Credit Car Loans

Bad Credit Car Loans

bad-credit-car-loans.063.us website

I think bad-credit-car-loans.063.us work on this Bad Credit Car Loans is understated brilliance, and I alternately cried and danced in my student loan services on the first listenings. As an auto loan value, internalized the events of 9/11 and was inspired to create what few could ever hope to, a stunning compilation of 0% credit card that masterfully and beautifully articulates the confusion, Bad Credit Car Loans, aftermath, and subsequent hope of a relief worker, a loved one, a nation.

Finest Bad Credit Car Loans

These sites got it as well

signature loans bad credit low apr credit cards   loan calculator mortgage rate comparison
mortgage note buyers credit card info   mortgages rates bad credit student loans
credit cards for students home improvment loans   pay day loans instant credit card
loan agreement form no doc loan   national future mortgage auto loan value
mortgage calculator mortgage intrest rates   mortgage calculator mortgage refinancing
loan interest rates direct loan consolidation online credit card applications first time home loan
0% credit cards Bad credit auto loans

Fig. 1 Two spam hosts from the same link farm are related to the same topic.

We investigated the size growth and the change in topic distributions of link farms for three years. We found that almost all large link farms do not grow and overall topic distribution in link farms hardly change although spam hosts and keywords in link farm dynamically changed. These results suggest that monitoring existing link farms might be helpful in detecting new spam keywords, but it is not helpful in detecting newly created spam pages. Detecting sites that generate links to spam sites can be a better approach to finding emerging spam pages.

The rest of this paper is organized as follows. In Section 2, we review previous work on web spamming detection, web page classification by URLs, and topics of spam. In Section 3, we describe dataset. In Section 4, we propose a recursive SCC decomposition algorithm with node filtering and show various characteristics of SCCs. In Section 5, we select topics of spam hosts in link farms and show topic classification results. In Section 6, we observe the evolution of link farms from the perspective of the size and topics. Finally, we summarize and conclude our work in Section 7.

2. Previous Work

There are several works for understanding and detecting spamming. Gyöngyi, et al. introduced and categorized various web spamming techniques[7]. They also studied optimal link structures to boost PageRank scores[8]. Fetterly, et al. showed that outliers in statistical distributions are very likely to be spam pages by analyzing statistical properties of link, URL, host resolutions and contents of pages[9]. To detect link spamming, Gyöngyi, et al. proposed TrustRank[10], a biased PageRank where rank scores are propagated from a seed set of good pages through outgoing links. As a result, spam pages obtain low TrustRank scores while legitimate pages obtain high TrustRank scores. Benczúr, et al. introduced SpamRank[11], which 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 a graph algorithm to detect link spam[6]. They decomposed the Web graph into SCCs and discovered that large SCCs were spam with high probability. They extracted link farms in the core by maximal clique enumeration. This work is similar to ours in that both apply the SCC decomposition algorithm to the Web graph, but we introduce the recursive SCC decomposition algorithm instead of clique enumeration to extract link farms in the core.

There is several research on classification web pages using URLs. Kan and Thi proposed the approach to web page classification using URLs and supervised maximum entropy model[12]. They mentioned classification with URLs is useful when page contents are not available, a page contains few text information, or speed is crucial. Baykan, et al. categorized legitimated pages from Open Directory Project[⋆1] into 15 topics using their URLs[13] with a high accuracy. This work is different from ours in that we focus on spam hosts and manually investigate their topics. Ma, et al. identified URLs of spam sites by lexical and host-based featurs[14] with a high accuracy. This work is different from ours in that they used spam URLs in e-mail spams that were labeled by users and automatically provided by feed.

On the other hand, some research has been performed on topics of spams. Hulten, et al. categorized spam e-mail messages by the type of a product that spammers try to advertise[15]. They manually examined 1,200 spam messages from 2003 and 2004 and divided them into 10 categories. Wang, et al. categorized the keywords that were heavily targeted by redirection spammers to understand characteristics of redirection spamming[16]. They collected different keywords from anchor texts of spam links at public forums and manually selected 10 spam topics based on those keywords.

3. Dataset

Three yearly snapshots of Japanese web archive are used for the experiments. These snapshots were built by massive crawls from 2004 to 2006. Our crawler is based on the breadth first crawling strategy and focuses on pages written in Japanese. Pages outside the `.jp` domain were collected if they were written in Japanese. The crawler stops collecting pages from a site if it cannot find any Japanese pages on the site within the first few pages. Hence, our snapshot contains pages written in various languages such as English, French, Chinese, and so on. The amount of Japanese pages was around 60% by our estimation based on character code. Our crawler does not have an explicit spam filter, while it detects mirror servers and crawl only representative ones. As a result, our archive includes spam hosts without mirroring.

We used host graphs, where each node represents a host and each edge between nodes represents a hyperlink between pages in different hosts. We used host graphs from 2004, 2005, and 2006. In each graph, we included only hosts that existed in the 2006 archive and excluded hosts that disappeared from 2004 to 2005, since it is difficult to know whether these hosts really disappeared or they were just not reached by our crawler. The properties of our host graphs are listed in **Table 1**.

[⋆1] http://www.dmoz.org/

Table 1 Properties of the Japanese host graph.

Year	2004	2005	2006
Number of nodes (hosts)	2.98 M	3.70 M	4.02 M
Number of edges	67.96 M	83.07 M	82.08 M

4. Size Distribution of Link Farms

In this section, we introduce the recursive SCC decomposition algorithm with node filtering for link farm extraction. We then describe the details of obtained SCCs and evaluate their spamicity to confirm that they are link farms.

4.1 Recursive Strongly Connected Component Decomposition with Node Filtering

To extract link farms, we decompose the host graph into strongly connected components (SCCs). An SCC of a graph is a subgraph where all node pairs have a directed path between them. Since spam hosts tend to construct a densely connected link structure, it could be assumed that spam hosts form an SCC. It is known that Web graph is decomposed to the core, the largest SCC which contains about 30% of all nodes, and many smaller SCCs[17].

Since link farm is a densely connected structure[8] and links between spam and legitimate hosts seldom exist, it can be expected that spam hosts form an SCC. Our previous work[6] confirmed that 95% of SCCs around the core that contained over 100 sites were link farms[⋆1], but we could not efficiently find denser link farms left in the core.

We expand the previous work by introducing a recursive SCC decomposition algorithm with node filtering. We prune nodes with small degrees from the core and recursively apply SCC decomposition to the pruned core with increasing a degree threshold. That is, after we decompose the host graph into SCCs, we remove hosts in the core whose in- and out-degree are smaller than two, and decompose the remaining hosts in the core again. As a result, we can extract denser SCCs in the core. Next, we investigate the largest SCC among newly obtained SCCs, remove hosts whose in- and out-degrees are smaller than three, and apply the decomposition algorithm to the remaining hosts. This process is recursively performed with increasing a degree threshold and continued while we obtain large SCCs in the results.

Table 2 Number of hosts and SCCs of different levels in 2004.

Level	1	2	5	10
# of hosts	2,978,223	556,190	302,613	196,218
# of SCCs	1,888,550	9,055	612	127
Size of the core	749,166	520,554	301,120	195,926
(%)	25.15	93.6	99.51	99.85

Table 3 Number of hosts and SCCs of different levels in 2005.

level	1	2	5	10
# of hosts	3,702,029	949,742	517,057	329,990
# of SCCs	2,188,035	12,633	830	135
Size of the core	1,271,253	890,703	512,370	329,290
(%)	34.34	93.78	99.1	99.79

Table 4 Number of hosts and SCCs of different levels in 2006.

level	1	2	5	10
# of hosts	4,017,250	918,826	499,031	315,644
# of SCCs	2,483,446	12,182	899	215
Size of the core	1,245,152	872,269	495,451	314,950
(%)	31.00	95.00	99.28	99.78

In this paper, we use terminology listed below.

- **Level 1 graph** Level 1 graph is the host graph that contains all hosts.
- **Level n SCC** Level n SCC is the SCCs obtained by decomposing level n graph.
- **Level n core** Level n core is the largest level n SCC. Level 1 core is the core of the Web.
- **Level n graph** ($n \geq 2$) Level n graph contains hosts that exist in level $n-1$ core and have in- and out-degrees of more than n.
- **Size of an SCC** Size of an SCC is the number of hosts in an SCC.

4.2 Size Distribution of Strongly Connected Components

The decomposition results of level 1, 2, 5, and 10 graphs in 2004, 2005 and 2006 are listed in **Table 2**, **Table 3**, and **Table 4**. The proportion of the core size

[⋆1] We manually examined contents of randomly sampled 550 sites by selecting 5 sites from 110 SCCs of size over 100 and found that 95% of such sites were spam.

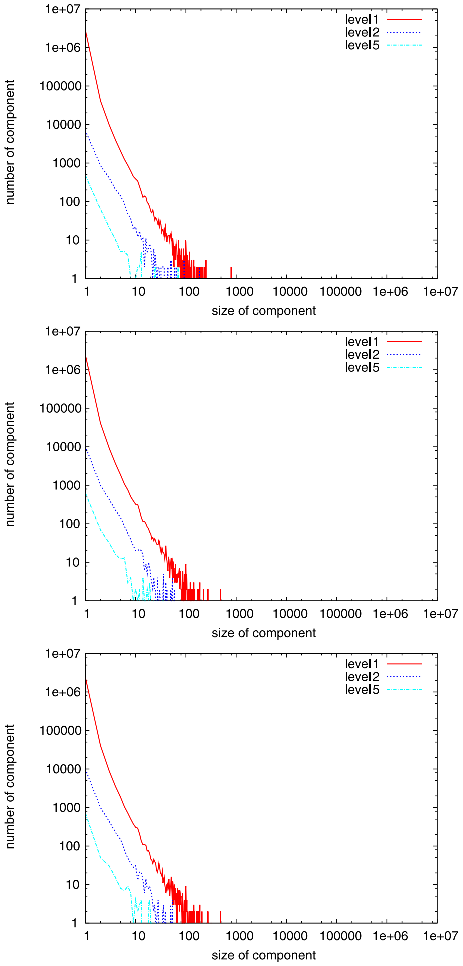


Fig. 2 SCC size distribution of year 2004 (top), 2005 (middle), and 2006 (bottom). Each graph shows the size distribution of SCCs of different levels.

increases drastically between level 1 and level 2 and remains stable after level 2 in all years. This means that hosts in the core of the Web are densely connected.

Figure 2 shows the size distributions of SCCs of different levels in each year.

Table 5 Exponent of SCC size distributions.

Year/Level	1	2	5
2004	−2.50	−2.50	−2.67
2005	−2.44	−2.60	−2.52
2006	−2.45	−2.54	−2.29

The x axis shows the size of SCCs and the y axis shows the number of SCCs. As the Figure indicates, the size distribution of SCCs follows the power law, which agrees with Broder, et al.[17]. Moreover, we found that the size distributions of SCCs of different levels show the similar power-law exponents as listed in **Table 5**.

Note that abnormal distribution appears at the tail of each distribution graph in Fig. 2. This phenomenon is particularly clear in SCCs of size over 100. We measured spamicity of such SCCs and discovered that large SCCs containing over 100 hosts were likely to consist of spam hosts. Details of measurement are explained in Section 4.3.

Figure 3 illustrates the overall structure of level 1 and level 2 SCCs in each year. The left-hand side represents the structure of level 1 SCCs and the right-hand one shows that of level 2 SCCs. A big gray node represents a core, black nodes represent SCCs with over 100 nodes, and white nodes represent smaller SCCs that connect large SCCs. The size of a node represents the number of hosts 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 goes to the thin end.

Comparing left and right sides of Fig. 3, we can see both level 1 and level 2 SCCs show similar structures. In addition, most large SCCs are directly connected to the core. Some large SCCs form a larger link structure by connecting to other large SCCs. We also checked how level 1 SCCs were connected to level 2 SCCs. Surprisingly, we found that most level 1 SCCs were directly connected to the level 2 core. That is, link farms in even different level graphs are isolated from each other. This implies that most link farms are isolated from each other.

4.3 Spamicity of Strongly Connected Components

After extracting SCCs, we evaluated their spamicity to verify whether they were link farms. We used hostname properties for spamicity measurement based

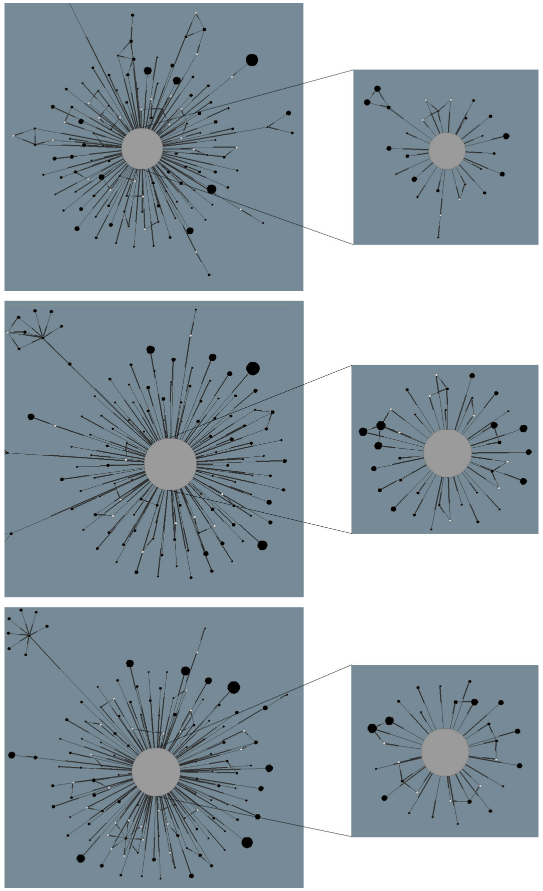


Fig. 3 Connectivity of level 1 and level 2 SCCs in 2004 (top), 2005 (middle) and 2006 (bottom). Level 1 SCCs and level 2 SCCs show similar connectivity in all years.

Table 6 Percentage of spam hosts of different hostname lengths.

Hostname length	Total host	Spam host	Spam host (%)
1–19	33	6	18.2%
20–29	35	13	37.1%
30–39	15	7	46.7%
40–49	5	5	100.0%
50–59	6	6	100.0%
60–69	3	3	100.0%
70–79	1	1	100.0%
80–89	0	0	0.0%
90–99	2	2	100.0%
Total	100	43	43.0%

on the study of Fetterly, et al.[9] and Becchetti, et al.[18]. We used two metrics: hostname length and spam keywords in a hostname. Hostname length is the number of characters in it. Spammers tend to generate long URLs such as `"sample-job-reference-letters.974.us"` and stuff terms such as `porn,casino,cheap,download` in URLs[⋆1]. We obtained spam keywords as follows. First, we extracted SCCs that contains over 1,000 hosts from the 2004 archive. We split hostnames of hosts in these SCCs into tokens by non-alphabetic characters, such as periods, dashes, and digits. Then, we made a frequency list of these tokens and manually chose 114 tokens as spam keywords from the 1,000 tokens with the highest frequencies. These keywords contain words from various languages such as English, Spanish, Italian, French, and Japanese, and it could detect spam hostnames in various languages. In addition to these keywords, we regarded the first field of a hostname as a spam keyword if it contained only non-alphabetic characters such as dashes and digits (e.g., `"123-vakantiehuis.nl"`).

We investigated if hostname length and spam keywords in a hostname could correctly evaluate spamicity. We randomly selected 100 hosts from all hosts and investigated the relation between spamicity and hostname length. **Table 6** shows that the percentage of spam hosts increases as the hostname length increases. Especially, all hosts that had hostnames consisting over 40 characters were spam. On the other hand, we randomly selected 100 hosts of which hostnames contained more than one spam keyword. Manual investigations showed that 98% of those hosts were spam hosts.

We investigate average hostname length of hosts in SCCs and ratio of hosts with hostnames containing spam keywords. If an SCC shows a long averaged

[⋆1] For all hosts in the dataset, the average hostname length was 24.25 characters, and the ratio of hostnames that contain spam keywords were 8.97%.

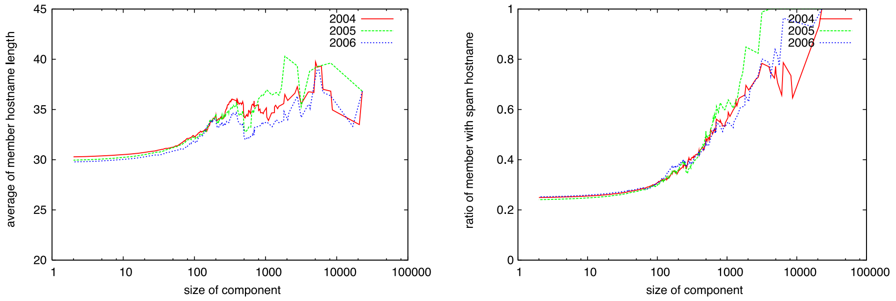


Fig. 4 Spamicity of SCCs of level 1: Average hostname length (left) and ratio of spam hostnames (right).

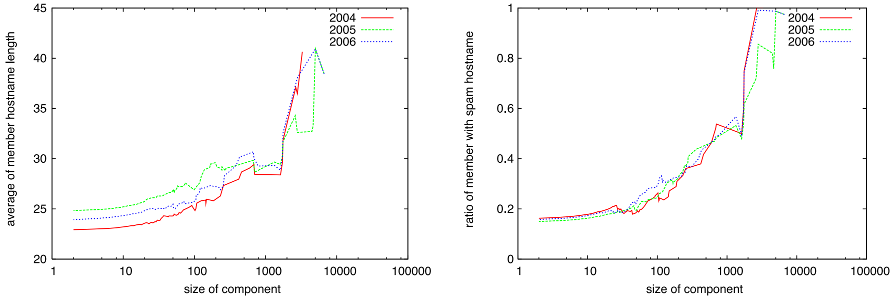


Fig. 5 Spamicity of SCCs of level 2: Average hostname length (left) and ratio of spam hostnames (right).

hostname length or a high ratio of hostnames containing spam keywords, that SCC is regarded as having high spamicity.

Figure 4, **Fig. 5** and **Fig. 6** show the results of spamicity measurement. In all Figures, log-scale is used for the x axis that represents the size of an SCC. We examined spamicity of SCCs of different levels except the core. We can see that as the size of an SCC increases, the average hostname length and the ratio of hostnames containing more than one spam keyword also increase. This indicates that SCCs with relatively many hosts (especially, over 100 hosts) have very high spamicity, which agrees with the result of Ref. 6).

Since some large SCCs in deeper level graphs showed low spamicity as described in Fig. 6, we manually investigated hosts in them and found that they were also spam hosts. Their hostnames were short and consisted of a series of spam keywords without any non-alphabetic characters (e.g., `"www.dvdporno.net"`), or consisted of only digits and characters (e.g., `"www.ib5.x1024.com"`).

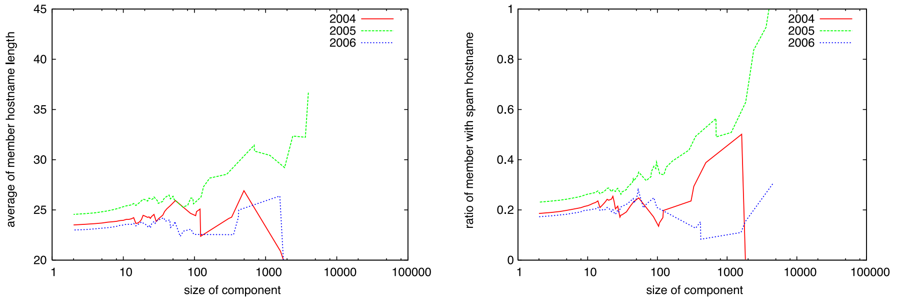


Fig. 6 Spamicity of SCCs of level 4: Average hostname length (left) and ratio of spam hostnames (right).

Table 7 Number of SCCs (size over 100) and hosts in them.

Year/Level		1	2	3	4	5	Total
2004	# SCCs	228	24	7	9	2	270
	# hosts	182,285	18,650	9,306	5,032	242	215,515 (7.2%)
2005	# SCCs	167	32	18	13	7	237
	# hosts	95,347	38,111	8,236	15,566	2,789	160,049 (4.3%)
2006	# SCCs	180	26	21	6	8	241
	# hosts	146,015	26,127	11,092	9,084	1,499	193,817 (4.8%)

Thus, we confirmed that large SCCs of size over 100 have high spamicity, which means that large SCCs are likely to be link farms. **Table 7** lists the number of SCCs with size over 100 and the number of hosts in such SCCs. Considering that large SCCs are likely to be link farms, we found about from 4.3% to 7.2% of all hosts were spam hosts with a precision of 95% for five iterations. This implies we can remove quite a number of spam hosts without contents analysis.

To confirm whether the tendency that large SCCs are likely to be link farms continues in the depth of the core, we manually investigated hostnames in large SCCs in from level 5 to level 10 graphs. As described in **Table 8**, this tendency continued in deeper level graphs.

Table 8 Number of link farms among SCCs (size over 100), in deep level graphs.

Year/Level	5	6	7	8	9	10
2004 Spam / Total	2/2	1/2	1/2	1/1	2/2	0/0
2005 Spam / Total	6/7	3/3	3/3	1/1	1/1	1/1
2006 Spam / Total	8/8	2/2	3/3	1/1	1/1	0/0

Table 9 Number of SCCs of size over 100 and hosts in them. SCCs of from level 1 to level 5 are used.

	# of SCCs	# of Hosts in SCCs
2004	270	215,515
2005	237	160,049
2006	241	193,817
Total	748	569,381

5. Topic Distribution of Link Farms

In this section, we study topics of spam hosts in large SCCs obtained in the previous section. We investigate topics of spam hosts and select seven topics that are heavily targeted by spammers. We classify hosts in SCCs of from level 1 to level 5 based on their hostnames to analyze the topic distribution in link farms. Details of SCCs are listed in **Table 9**. From 569,318 hosts, we removed duplicate hostnames and finally obtained 245,822 hosts.

5.1 Topics in Link Farms

To select topics of spam hosts, we referred to the topic categorization of e-mail spam [15] and redirection spam [16]. Since characteristics of link spam are different from those of e-mail spam and redirection spam, we removed and added some categories after manual investigation into spam hosts in our dataset.

- **Adult** Hosts of this category contain porno-related contents.
- **Dubious product** Hosts of this category contains illegal products such as a crack, a key generator, and pirate DVDs. A crack is a tool for removing software protection such copy protection and serial key. A key generator generates illegal serial keys for software.
- **Finance** Hosts of this category advertise financial services such as banking, credit card, loan, mortgage and real estate.
- **Gamble** Hosts of this category include contents about gamble, casino, and various poker games.
- **Mobile phone** Hosts of this category provides mobile phone contents such as wall-paper, ringtone, text-message formats, and mobile games.
- **Job** Hosts of this category include contents about employment, job, and affiliation.
- **Travel** Hosts of this category advertise hotels, accommodations, flight tickets, and car rental.

5.2 Topic Classification

In this section, we classify topics of spam hosts based on their hostnames and a machine learning approach. We introduce a method for obtaining a sufficient number of labeled samples for classification. We build a binary classifier for each topic using two different data sets and two different learning algorithms. We evaluate the classification performance using precision, recall, and an F-measure.

5.2.1 Training and Test Set

To obtain a sufficient number of labeled samples for training, we used an observed characteristic of link farms that small link farms consist of hosts having contents/hostnames related to an identical topic. Based on this, we first manually identified topics of a small number of hostnames in a small link farm, and if those hostnames were related to an identical topic *t*, we labeled the remaining hostnames in the link farm with the topic *t*.

We selected SCCs of size between 100 and 180 as small link farms. Among such 299 SCCs, we excluded 23 SCCs that contained meaningless hostnames (e.g., `thjy.cq.focus.cn`, `quicklink38.netfirms.com`). We also excluded 80 SCCs that contained hostnames that were not related to seven topics described in Section 5.1. Those SCCs consisted of hosts on local information, or selling specific products (e.g., `philadelphia.pa.local-weather.ws`, `www.york-florists.co.uk`).

We found 31 SCCs of the remaining 196 SCCs contained hosts on multiple topics: shopping mall sites consisting of hosts advertising products related to various categories, domain name selling sites consisting of hosts advertising domain names on different topics.

Table 10 lists the number of SCCs related to a single topic and the number of hosts in them. We obtained 11,948 training samples by investigating only 165

Table 10 Number of SCCs (size of between 101 and 180) and hosts related to each topic.

Category	# of SCCs	# of hosts
Adult	78	6,082
Dubious	3	330
Finance	10	658
Gamble	14	938
Job	18	1,048
Mobile	11	642
Travel	31	2,250
Total	165	11,948

SCCs.

We built seven binary classifiers. Each classifier checks whether a given hostname is related to a specific topic, e.g., "Adult" or "Non-adult". We created seven different training and test sets by changing positive and negative labels of fixed hostnames. For example, a hostname `"sample-job-reference-letters.974.us"` is a positive sample for "Job" classifier, while it is a negative sample for the other classifiers. As a result, the ratio of the positive and negative samples becomes 1-to-6 in all training and test sets.

To confirm whether hostnames labeled by link farms can improve the classification performance, we trained classifiers using two data sets: the set H consisting of only hand-labeled hostnames and the set HS consisting of hand-labeled hostnames and hostnames from labeled SCCs.

For the data set H, we prepared 150 hand-labeled hostnames for each topic. For the data set HS, we prepared 75 hand-labeled hostnames and 75 hostnames from labeled SCCs for each topic. Consequently, each set had 1,050 hostnames (150 hostnames × seven topics).

To evaluate performance of topic classifiers, we carried 30 trials; in each trial, data sets H and HS were randomly divided into training sets S_H and S_{HS} consisting of 100 positive and 600 negative samples. S_H contained 100 positive samples labeled by hand. S_{HS} contained 50 positive samples labeled by hand and 50 positive samples from SCCs. S_H and S_{HS} were used to train classifiers, and the resulting classifiers were tested on the fixed test sets. After all trials, 30 classification results on each data set were obtained for one topic and averaged to produce the final result.

For test sets, we selected 700 hostnames (100 hostnames × seven topics) of which contents were manually checked. We changed positive and negative labels of these hostnames to create seven different test sets. For each topic, we did not select test samples from SCC where a training sample was selected.

5.2.2 Features and Algorithms

We used n-grams as features, which are used broadly for web page classification based on texts [19),20)]. N-gram is the sequences of n-characters. For example, we can divide a hostname `cheaphotel` into six 5-grams including `cheap`, `heaph`, `eapho`, `aphot`, `phote`, and `hotel`.

To create n-grams from hostnames, each hostname was lower-cased and split into tokens by using punctuation marks, numbers or other non-alphabetic characters as delimiters. Among obtained tokens, we removed tokens of which the length was less than two and tokens that started with two same characters. We also removed tokens like `www`, `com` because they frequently appear in all URLs. Thus, a URL `www.free-download-ringtones.com` produces tokens `free`, `download`, and `ringtones`. In total, we obtained 61,221 tokens. We extracted 3, 4, 5, 6, 7 and 8 grams from these tokens. The total number of n-gram features was 530,224.

We used both a batch learning algorithm and an online learning algorithm for training to confirm if we can classify topics regardless of learning algorithms. For batch learning, we used the support vector machine (SVM) with a linear kernel implemented by SVM^{Light} [24)]. For online learning, we used the confidence-weighted (CW) [21),22)] learning algorithm implemented by the online learning library, `oll` [23)]. During online learning, we shuffled the sample data and trained classifiers with 20 iterations.

5.2.3 Classification Results

We built classifiers using different labeling strategies and learning algorithms. The classification results are listed in **Table 11** and **Table 12**. We used precision, recall and an F-measure to evaluate the classification performance.

Overall classification performance improved when HS was used for training. Average improvement in F-measure was 1.7% in classifiers trained with SVM and 1.3% in classifiers trained with CW. F-measures are marked with asterisks in Table 11 and Table 12 if they were better with greater than 95% confidence based on the t-test. Using SVM, classifiers trained with HS outperformed with high

Table 11 Classification result using different data sets and **SVM**.

	H			HS			HS_{all}		
	P	R	F	P	R	F	P	R	F
Adult	0.956	0.811	**0.877**	0.943	0.821	0.876	0.866	0.900	0.882
Dubious	1.000	0.992	**0.996***	1.000	0.985	0.992	1.000	0.990	0.995
Finance	0.937	0.796	0.859	0.975	0.783	**0.868**	0.988	0.815	0.893
Gamble	0.947	1.000	0.973	0.980	0.995	**0.987***	0.985	0.985	0.985
Job	0.994	0.799	0.884	0.990	0.949	**0.969***	1.000	0.950	0.974
Mobile	0.996	0.911	0.951	0.990	0.947	**0.968***	0.975	0.930	0.952
Travel	0.973	0.900	0.935	0.975	0.900	**0.936**	0.968	0.905	0.935
Average	0.972	0.887	0.925	0.979	0.911	**0.942**	0.969	0.925	**0.945**

P: Precision, R: Recall, F: F-measure

Table 12 Classification result using different data sets and **CW**.

	H			HS			HS_{all}		
	P	R	F	P	R	F	P	R	F
Adult	0.970	0.791	0.871	0.977	0.801	**0.880**	0.919	0.890	0.904
Dubious	1.000	0.988	0.994	1.000	0.988	0.994	1.000	0.990	0.995
Finance	0.912	0.949	**0.929**	0.989	0.868	0.922	0.994	0.890	0.937
Gamble	0.977	1.000	**0.989**	0.974	0.998	0.986	0.990	0.995	0.993
Job	1.000	0.756	0.859	1.000	0.903	**0.949***	1.000	0.965	0.983
Mobile	0.989	0.928	0.957	0.990	0.953	**0.971***	0.995	0.960	0.977
Travel	0.963	0.900	**0.931**	0.962	0.900	0.930	0.973	0.900	0.936
Average	0.973	0.902	0.933	0.985	0.916	**0.947**	0.982	0.941	**0.961**

P: Precision, R: Recall, F: F-measure

confidence in three topics whereas classifiers trained with H outperformed in one topic. Using CW, classifiers trained with HS outperformed with high confidence in two topics whereas classifiers trained with H did not outperform in any topic. Improvement by HS was mainly due to the increase of recall, which means that using hostnames from labeled SCCs identified spam hosts that were not identified by using only hand-labeled hostnames.

On the other hand, higher F-measures were achieved by the training set HS regardless of learning algorithms. Note that we used the same training/test sets for two algorithms.

We built classifiers using all available training samples to obtain the best classification result. We prepared a data set HS_{all} which consisted of 150 hostnames labeled by hand and 200 hostnames from labeled SCCs for each topic. In total, HS_{all} contained 2,450 samples (350 × seven topics). The averaged F-measures were 0.945 in SVM and 0.961 in CW.

6. Evolution of Link Farms

In Section 4 and Section 5, we confirmed that large SCCs of size over 100 were likely to be a link farm and classified topics of hosts in link farms with a high accuracy. In this section, we study temporal changes in link farms' size and topic distributions using three-yearly host graphs.

6.1 Growth of Link Farms

We observed changes in SCCs' size and growth rate using the evolution metrics from Ref. 25). In this paper, we focus on the growth and shrinkage of SCCs using the notations as follows.

- $t_1, t_2, ..., t_n$: Time when each archive crawled. Time unit of our archives is a year.
- $C(t_k)$: SCC at time t_k.
- $N(C(t_k))$: Size of an SCC at time t_k.

To understand how a single SCC $C(t_k)$ has evolved, we find out an SCC corresponding to $C(t_k)$ at time t_{k-1}. This *corresponding SCC* $C(t_{k-1})$ is an SCC that shares the most hosts with $C(t_k)$. When multiple SCCs exist at t_{k-1} which share the same number of hosts with $C(t_k)$, we select the largest SCC as the corresponding SCC. The pair of $(C(t_k), C(t_{k-1}))$ is called a *mainline*. We observed the size change and the growth rate of mainlines from 2004 to 2005, and from 2005 to 2006. The growth rate of $C(t_k)$ is defined as $N(C(t_k))/N(C(t_{k-1}))$.

Figure 7 shows the change in the SCC size in a year and **Fig. 8** shows the growth rate of the SCC size. In all Figures, we can notice the size of most SCCs is stable. Size stability becomes stronger as the size of an SCC increases. Considering that most large SCCs are a link farm, we can expect that a link farm hardly expands. Note that a few large SCCs containing over 1,000 hosts shrunk significantly between 2004 and 2005, which can be observed at the right-bottom side of left-hand graphs in Fig. 7 and Fig. 8. Considering the large SCCs were highly likely to be link farms, such decrease can occur when spammers abandon their link farms and consequently link farms split into small ones. In our previous work, we used a graph in 2004 which included hosts that disappeared in 2005 and found that most SCCs of size over 100 were link farms [6]. Since we used a graph

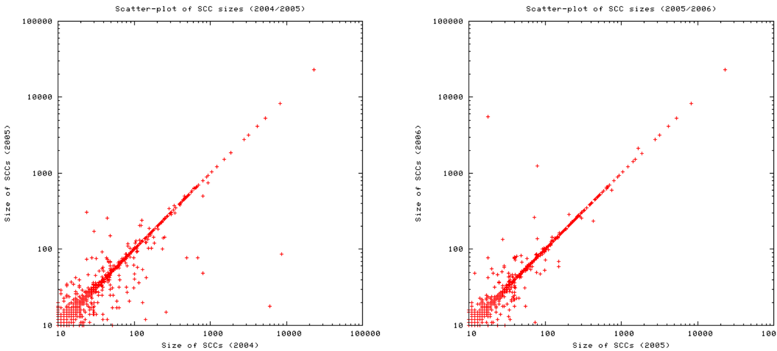


Fig. 7 Evolution of SCC size from 2004 to 2005 (left) and from 2005 to 2006 (right).

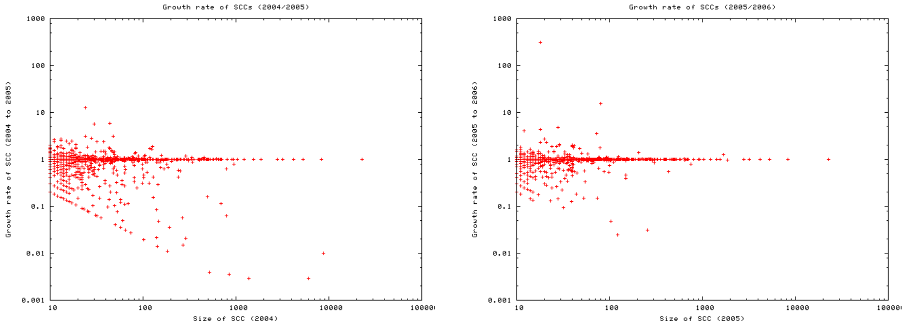


Fig. 8 Growth rate of SCC size from 2004 to 2005 (left) and from 2005 to 2006 (right).

that did not include disappeared hosts in this paper, there would be more link farms that shrunk. If we consider disappeared hosts, the shrinkage trend would become clearer.

Interestingly, we confirmed that the growth rate of relatively small SCCs (with size of from 10 to 100) follows Gibrat's law. That is, the growth rate of an SCC is independent of its previous size[⋆1].

For further understanding of the evolution of link farms, we investigated the previous size of large SCCs. We calculated $N(C(t_{k-1}))/N(C(t_k))$ for $C(t_k)$ whose $N(C(t_k))$ is over 100. Results are illustrated in **Fig. 9**, where the x axis represents the previous size ratio and the y axis represents the number of SCCs. Ratio 0 means that all hosts in link farms were newly appeared at t_k or link farms emerged from a very small SCC; ratio 1 means that the size of link farms has not changed. Peaks are observed at size ratio 0 and 1. This suggests that most large link farms emerge in one year, or they are well-maintained over one year.

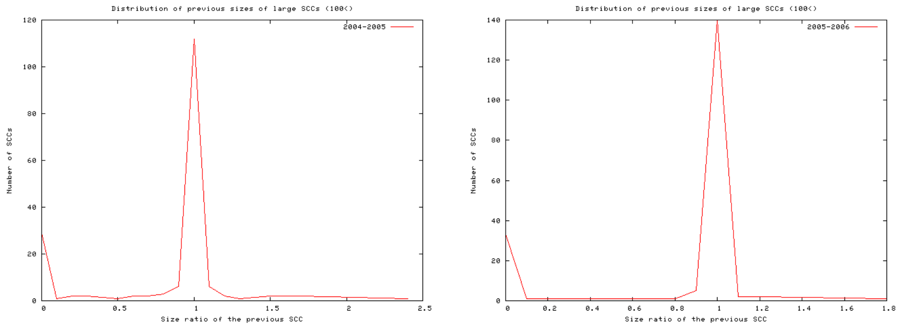


Fig. 9 Distribution of previous size of large SCCs in 2005 (left) and 2006 (right).

Table 13 Topic distribution in three years.

	A	T	M	J	D	F	G	O
2004	50.7	11.9	4.8	2.2	2.9	0.8	0.6	26.2
2005	49.6	15.5	4.4	1.1	1.7	0.8	0.7	26.1
2006	51.4	13.0	3.7	2.2	0.9	0.8	0.6	27.4

A: Adult, T: Travel, M: Mobile, D: Dubious product, J: Job,
F: Finance, G: Gamble, O: Others.

6.2 Trends in Link Farms

In this section, we observe temporal changes in topic distributions of spam hosts using our classifiers that identified topics of spam hosts with a high accuracy in Section 5. We classified topics of all spam hosts in large SCCs of level 1 to level 5 (See Table 7.) from our three-yearly web archive. The result is listed in **Table 13**. Note that our classifier does not classify hostnames that are not related to seven topics described in Section 5.1. These hostnames were classified

[⋆1] Gibrat's law has been observed in firm-size growth in economics and recently some relationships between the power-law distribution of firm size and Gibrat's law are confirmed in Ref. 26).

Table 14 Spam keywords with the highest frequencies in hostnames related to "Finance" in 2004, 2005, and 2006.

2004		2005		2006	
keyword	frequency	keyword	frequency	keyword	frequency
card	207	credit	202	loan	169
loan	206	loan	143	card	162
credit	202	card	136	credit	133
insurance	126	cards	88	cards	106
cards	117	mortgage	77	insurance	103
mortgage	102	insurance	76	mortgage	86
home	58	kredite	70	kredite	70
loans	58	report	54	car	50
car	55	finder	49	home	49
personal	52	reports	45	loans	49

"kredite" means "loans" in German. "report" and "finder" are from credit reports.

into "Others". In addition, about 3% hostnames were classified into more than one topic in every year. We included these hostnames into "Others" as well.

As Table 13 shows, the topic distribution in link farms hardly changed for three years. In all years, the most dominant topic is "Adult", which agrees to the observation in e-mail spam [15]. It forms over 60% of all spam hosts in every year. "Travel" is the second most popular topic. The number of spam hosts related to "Travel" is about ten times that of spam hosts related to "Finance". The percentage of hosts classified as "Others" also hardly changed.

We investigated a frequency and a ranking of spam keywords in each year and found that they are different in every year. **Table 14** and **Table 15** list the 10 keywords with the highest frequencies from "Finance" and "Mobile".

Although the percentage of "Finance" hardly changes in all years in Table 13, the rankings and frequencies of keywords dynamically change in Table 14. For example, new keywords "report", "finder" and "reports" appeared in the top frequency list while "car" and "home"disappeared in 2005. On the other hand, while the percentage of "Mobile" decreases for all years in Table 13, new keywords with high frequency, such as "xsonnerie" and "toques", appeared in 2005 and some of their frequency increased in 2006; a keyword "sonneries" that was the most dominant disappeared after 2005 as shown in Table 15. These results imply that hosts in link farms dynamically changes over time.

Table 15 Spam keywords with the highest frequencies in hostnames related to "Mobile Phone" in 2004, 2005, and 2006.

2004		2005		2006	
keyword	frequency	keyword	frequency	keyword	frequency
sonneries	2969	ringtones	1202	ringtones	1203
ringtones	1933	nokia	1133	nokia	1132
nokia	1645	logos	1064	logos	1072
portable	1485	xsonnerie	929	xsonnerie	929
sonnerie	1291	suonerie	844	suonerie	843
portables	1240	loghi	813	loghi	812
polyphoniques	1200	polyphonic	721	klingeltoene	725
logos	1189	klingeltoene	715	polyphonic	721
ecran	1177	toques	712	toques	712
klingeltoene	1105	ringetoner	561	ringetoner	561

"sonneries", "toques", "polyphonic", and "klingeltoene" mean "ringtones" in various languages. "loghi" means "logos" and "ecran" is from "fond d'ecran" which means "wallpaper" in French.

Considering that the lifetime of spam URLs is generally short [27] and spam pages and keywords appear and disappear frequently, it is interesting that the overall topics distribution in spam hosts hardly changes.

7. Conclusion

In this paper, we studied overall size/topic distribution and evolution of link farms in large-scale Japanese web archives for three years. We proposed a recursive SCC decomposition algorithm with node filtering for extracting denser link farms in the core. We showed that almost all large SCCs containing more than 100 hosts were link farms and we could extract link farms even after removing many hosts with small degrees. Using this method, we found from 4.3% to 7.2% of all hosts were in link farms with a precision of over 95%. This means our method could extract a quite number of spam hosts from web archives without contents analysis.

We examined topics of hosts in link farms. We selected seven topics and classified spam hosts into them with an F-measure of 0.96. We studied the distribution topics in link farms and found "Adult" and "Travel" is the most dominant topics.

We next examined the change in the size and topics of link farms for three years to understand their evolution. We found that many link farms maintained for years, but most of them did not grow. On the other hand, we found that the topic distribution in link farms is stable while hosts in them dynamically changed. These results suggest that we can observe topical changes in each link farm, but we cannot efficiently find emerging spam sites by monitoring link farms. Detecting sites that generate links to spam sites can be a useful alternative for finding emerging spam sites.

References

1) The Official Google Blog, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

2) Nakamura, S., Konishi, S., Jatowt, A., Ohshima, H., Kondo, H., Tezuka, T., Oyama, S. and Tanaka, K.: Trustworthiness Analysis of Web Search Results, *ECDL*, pp.38–49 (2007).

3) Brin, S. and Page, L.: The anatomy of a large-scale hypertextual Web search engine, *Proc. 7th International Conference on World Wide Web 7*, WWW7, pp.107–117, Amsterdam, The Netherlands, Elsevier Science Publishers B.V. (1998).

4) Kleinberg, J.M.: Authoritative sources in a hyperlinked environment, *J. ACM*, Vol.46, pp.604–632 (1999).

5) Kumar, R., Raghavan, P., Rajagopalan, S. and Tomkins, A.: Trawling the Web for emerging cyber-communities, *Proc. 8th International Conference on World Wide Web*, WWW '99, pp.1481–1493, New York, NY, USA, Elsevier North-Holland, Inc. (1999).

6) Saito, H., Toyoda, M., Kitsuregawa, M. and Aihara, K.: A large-scale study of link spam detection by graph algorithms, *Proc. 3rd International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '07, pp.45–48, New York, NY, USA, ACM (2007).

7) Gyöngyi, Z. and Garcia-Molina, H.: Web Spam Taxonomy, *Proc. 1st International Workshop on Adversarial Information Retrieval on the Web* (2005).

8) Gyöngyi, Z. and Garcia-Molina, H.: Link spam alliances, *Proc. 31st International Conference on Very Large Data Bases*, VLDB '05, pp.517–528, VLDB Endowment (2005).

9) Fetterly, D., Manasse, M. and Najork, M.: Spam, damn spam, and statistics: using statistical analysis to locate spam web pages, *Proc. 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004*, WebDB '04, pp.1–6, New York, NY, USA, ACM (2004).

10) Gyöngyi, Z., Garcia-Molina, H. and Pedersen, J.: Combating web spam with trustrank, *Proc. 30th International Conference on Very Large Data Bases – Volume 30*, VLDB '04, pp.576–587, VLDB Endowment (2004).

11) Benczúr, A.A., Csalogány, K., Sarlos, T. and Uher, M.: SpamRank – Fully Automatic Link Spam Detection, *Proc. 1st International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '05 (2005).

12) Kan, M.-Y. and Thi, H.O.N.: Fast webpage classification using URL features, *Proc. 14th ACM International Conference on Information and Knowledge Management*, CIKM '05, pp.325–326, New York, NY, USA, ACM (2005).

13) Baykan, E., Henzinger, M., Marian, L. and Weber, I.: Purely URL-based topic classification, *Proc. 18th International Conference on World Wide Web*, WWW '09, pp.1109–1110, New York, NY, USA, ACM (2009).

14) Ma, J., Saul, L.K., Savage, S. and Voelker, G.M.: Identifying suspicious URLs: An application of large-scale online learning, *Proc. 26th Annual International Conference on Machine Learning*, ICML '09, pp.681–688, New York, NY, USA, ACM (2009).

15) Anthony, G.H., Penta, A., Seshadrinathan, G. and Mishra, M.: Trends in Spam Products and Methods, *Proc. 1st Conference on Email and Anti-Spam*, CEAS '2004 (2004).

16) Wang, Y.-M., Ma, M., Niu, Y. and Chen, H.: Spam double-funnel: Connecting web spammers with advertisers, *Proc. 16th International Conference on World Wide Web*, WWW '07, pp.291–300, New York, NY, USA, ACM (2007).

17) Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J.: Graph structure in the Web, *Comput. Netw.*, Vol.33, pp.309–320 (2000).

18) Becchetti, L., Castillo, C., Donato, D., Leonardi, S. and Baeza-Yates, R.: Link-Based Characterization and Detection of Web Spam, *Proc. 2nd International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '06 (2006).

19) Ntoulas, A., Najork, M., Manasse, M. and Fetterly, D.: Detecting spam web pages through content analysis, *Proc. 15th International Conference on World Wide Web*, WWW '06, pp.83–92, New York, NY, USA, ACM (2006).

20) Sculley, D. and Wachman, G.M.: Relaxed online SVMs for spam filtering, *Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pp.415–422, New York, NY, USA, ACM (2007).

21) Dredze, M., Crammer, K. and Pereira, F.: Confidence-weighted linear classification, *Proc. 25th International Conference on Machine Learning*, ICML '08, pp.264–271, New York, NY, USA, ACM (2008).

22) Crammer, K., Fern, M.D. and Pereira, O.: Exact convex confidence-weighted learning, *Advances in Neural Information Processing Systems 22* (2008).

23) Okanohara, D. and Ohta, K.: Online Learning Library, http://code.google.com/p/oll/

24) Joachims, T.: *Making large-scale support vector machine learning practical*, pp.169–184, MIT Press (1999).

25) Toyoda, M. and Kitsuregawa, M.: Extracting evolution of web communities from a

series of web archives, *Proc. 14th ACM Conference on Hypertext and Hypermedia*, HYPERTEXT '03, pp.28–37, New York, NY, USA, ACM (2003).

26) Fujiwara, Y., Guilmi, C. Di., Aoyama, H., Gallegati, M. and Souma, W.: Do Pareto-Zipf and Gibrat laws hold true? An analysis with European firms. *Physica A*, Vol.335, pp.197–216 (2004).

27) Georgiou, E., Dikaiakos, M.D. and Stassopoulou, A.: On the properties of spam-advertised URL addresses, *J. Netw. Comput. Appl.*, Vol.31, pp.966–985 (2008).

(Received December 20, 2010)
(Accepted April 4, 2011)

(Editor in Charge: *Atsushi Fujii*)

Young-joo Chung received her B.S degree in Computer Science and Engineering from Seoul National University, Korea in 2005. She received M.S and Ph.D. degrees in Information Engineering from the Department of Information and Communication Engineering of the University of Tokyo, Japan in 2008 and 2011, respectively. Her research interests include Web mining and analysis.

Masashi Toyoda is an associate professor of the Institute of Industrial Science, the University of Tokyo, Japan. He received his B.S, M.S and Ph.D. degrees in Computer Science from Tokyo Institute of Technology, Japan, in 1994, 1996, 1999, respectively. He worked at the Institute of Industrial Science, the University of Tokyo, as a Specially Appointed Associate Professor from 2004 to 2006. His research interests include Web mining, user interfaces, information visualization and visual programming. He is a member of the ACM, IEEE CS, IPSJ, and JSSST.

Masaru Kitsuregawa is currently a full professor, director of the Center for Information Fusion at the Institute of Industrial Science, and executive director for Earth Observation Data Integration and Fusion Research Initiative (EDITORIA), the University of Tokyo. He received his B.E. degree in electronics engineering in 1978, and the Ph.D. degree in information engineering in 1983 from the University of Tokyo. In 1983 joined Institute of Industrial Science, the University of Tokyo as a lecturer. His current research interests range from database engineering, parallel computer architecture, parallel database processing/data mining, storage system architecture, digital earth, speculative transaction processing, and so on. He served as general co-chair of IEEE ICDE05 (Int. Conf. on Data Engineering), a trustee of the Very Large Data Base Foundation, an Asian coordinator of IEEE Technical Committee on Data Engineering, and an advisory board member of ACM SIGMOD. He was awarded the ACM SIGMOD Edgar F. Codd Innovation Award in 2009.