A Topical Study on the Web Spam

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

In this paper, we study the topical characteristic of spam hosts. To categorize spam hosts, we extract link spam structures from multiple time snapshots of Japanese Web archive using graph algorithms. Next, we define several spam topic categories and classify spam hosts in those structures into such spam topics using their uniform resource locator(URL)s and a machine learning approach. We analyze the spam topic distribution on the Web in different years and observe the change in spam topics through the time.

1. Introduction

As the Web becomes a major source of information, the amount of pages related with commerce also increases. People access web to buy products, reserve their hotel and flight, and apply for jobs and mortgages. This leads to the emergence of malicious pages, or *web spam*, which tries to yield profit by attracting users with unfair ways. Link spamming is one of the spamming techniques that manipulate the link structure of the Web. Spammers create a densely connected link structure, or a *link farm*, to boost link-based rank scores and raise a ranking in the search result list [1].

In this paper, we categorize spam hosts in link farms into different topics. Although there are researches on the topic of the whole Web [2] and e-mail spam [3], as far as we know, topical study focusing on web spam has not been performed.

To categorize spam hosts, first, we extract link farms from the web by strongly connected component decomposition with node filtering. This approach decompose the whole web graph into strongly connected component(SCC)s, and then removes nodes with small degrees from the largest SCC. SCC decomposition is applied to the pruned core recursively with increasing node degree threshold [5].

After we obtain link farms, we build a topic classifier to identify the topic of each spam hosts in link farms. Our classifier use uniform resource locator(URL)s to generate features. Classification using URL is fast and safe because it does not need to download all contents.

We test our classifiers on the large scale Japanese web archive and divide spam hosts into several topics. Also, we observe the difference in spam topic distribution though the time.

2. Link Farm and Spam Topics

2.1 Link Farm and Strongly Connected Component

Link farm is a densely inter-connected link structure created by spammer to maximize PageRank score. The link farm consists of a target page and boosting pages. All boosting pages link to the target page in order to increase the rank score of it. Then, the target page distributes its boosted PageRank score back to supporter pages. By this, members of a spam farm can boost their PageRank scores.

To extract link farm in the Web, we proposed strongly connected component decomposition with node filtering. Strongly connected component (SCC) is a subgraph where every pair of nodes has a directed path between them. Since link farm is a densely connected link structure [1], and links between spam and normal site hardly exist, it can be expected that spam sites form a SCC [4]. We extracted SCCs after decomposing the whole nodes in the graph. Next, we filter out nodes in the core of which in-degree and out-degree are smaller than 2, and decompose the remaining hosts in the core into SCCs again. As a result, we extract denser SCCs in the core. Next, we consider the largest among newly obtained SCCs, and remove nodes of which in and out degrees are smaller than 3, and apply the decomposition algorithm to the remaining hosts. This process is performed recursively with incrementing degree threshold, and continued while we have large SCCs in the results.

To confirm whether a SCC is a link farm or not, we use two metric; hostname length and spam keyword ratio. Spammers tend to generate long URLs like **"sample-job-reference-letters.974.us"** and stuff terms like **porn, casino, cheap, download** in URLs. Our previous work confirmed that SCC of size over 100 is very likely to be a link farm [5]. Therefore, we regard a SCC of size over 100 as a link farm and a host in such SCC as spam.

2.2 Spam Topic Categorization based on URLs

To understand the topical characteristic of link farms, we consider the classification of spam hostnames. We categorize spam hostnames into 7 categorizes referring to e-mail spam categorization [3], and performing manual checks on randomly selected hostnames. Table 1 shows the detail.

Table 1 Topic categorization of Web spam

Topic	Description
Adult	Sexual, pornographic contents
Dubious	Illegal downloads of serial keys, pirate DVDs
Financial	Insurance, mortgage, credit card
Gamble	Casino, poker and other gaming contents
Jobs	Employment search, home affiliation
Mobile	Download mobile contents like ringtone, wall-paper
Travel	Hotel reservation, search, flight reservation

There are a number of hostnames that advertise a specific product or provide regional information such as weather, college information or news, but we do not define topic categories for them.

For classification, we take machine learning approach. Based on lexical features of URLs, we construct multiple binary classifiers for each topic. N-gram, a sequential of *n* letters, is used as a feature. Spammers deliberately misspelled words like **"cheaap", "m0rtgage"** to avoid spam filters and such spam keywords can be detected using *n*-gram features.

We build our classifiers by confidence weighted learning algorithm [6] implemented by [7]. As for training samples, in addition to hand labeled hosts, we use hostnames that are from categorized small link farms. It can be assumed that a small link farm contains spam hosts related with a single topic. By adding hostnames from SCCs, we can secure a number of training samples and improve the performance of classification. The experiment on hostnames showed high performance and average F-measure of all topic classification results was 0.996 [8]. Therefore, we use these classifiers to categorize spam hosts.

3. Experiment

3.1 DATASET

We extracted link farms from the large scale snapshots of Japanese Web archive. These snapshots are built by crawling that conducted from 2004 to 2006. 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. Host graphs for 2004, 2005 and 2006 were built. The properties of our host graphs are shown in Table 2.

Table 2 The properties of host graphs

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

3.2 Spam Hosts in Link Farms

From yearly host graphs, we extracted link farms. SCC decomposition with node filtering(See Section 2.1) was performed during five iteration and SCCs of which size is over 100 are extracted as a link farm. The number of SCCs of size over 100 and hosts in those SCC are described in Table 3.

Table 3 The number of SCCs of size over 100 and host in them. SCCs are obtained during 5 iterations.

	#of SCCs(100<)	# of Hosts in SCCs
2004	268	215,515
2005	235	160,049
2006	239	193,817
Total	742	569,381

3.3 The Topical Distribution of Spam Hosts on the Web

In order to understand general topical distribution of spam, we classify hostnames in SCCs from different years. Table 4 demonstrates the result. Note that there are hostnames that do not belong to any of seven topics described in Table 1.

Table 4 The percentage of hostnames related with an each topic.

	A	D	F	G	J	M	T
2004	58.92	1.04	1.36	0.82	2.36	5.54	13.06
2005	57.85	1.97	1.62	1.06	1.18	5.14	17.22
2006	60.90	1.37	1.38	1.08	2.33	4.20	14.10

In all years, most dominant topic is an adult related topic. It forms over 50% of all spam hosts in every year. It is also remarkable that travel related topic is second most popular. The number of spam hosts in travel category is about ten times as many as that of in financial category. This might imply that people search travel and flight information more frequently through the Web, because the Web provide global information as well as domestic one. Finance related information, such as mortgage or credit card, has rather local characteristics and can be provided by banks or related institutes.

The percentage of spam hosts that related with gamble increases continuously, while that of mobile related spam host are decreases.

4. Conclusion

In this paper, we analyzed the distribution of spam topics on the Web. For this, we extracted a spam link structure, or a link farm, from the time series of web snapshots. In addition to this, we defined spam topic categorizes and built a classifier for each topic. After classification, we found that most popular spam topic is adult related topic which is followed by travel related topic.

REFERENCES

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

[2] S. Chakrabarti, M. M. Joshi, K. Punera, and D. M. Pennock. "The Structure of Broad Topics on the Web, " Proc. the 11th international conference on World Wide Web, 2002.

[3] G. Hulten, A. Penta, G. Seshadrinathan, and M. Mishra, "Trends in Spam Products and Methods, " Proc. the First Conference on Email and Anti-Spam, 2004.

[4] 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.

[5] Y. Chung, M. Toyoda, and M. Kitsuregawa, "A study of link farm distribution and evolution using a time series of web snapshots, " Proc. the 5th International workshop on adversarial information retrieval on the Web, 2009.

[6] M. Dredze, K. Crammer, F. Pereira. "Confidence-Weighted Linear Classification," Proc. the 25th International Conference on Machine Learning, 2008.

[7] D.Okanohara and K. Ohta. Online Learning Library. http://code.google.com/p/oll/

[8] Y. Chung, M. Toyoda, and M. Kitsuregawa, "Topic Classification of Spam Host based on URLs," Proc. the 2nd Forum on Data Engineering and Information Management(DEIM '10), 2010.