Detecting Search Engine Spam from a Trackback Network in Blogspace

Masahiro Kimura[1], Kazumi Saito[2], Kazuhiro Kazama[3], and Shin-ya Sato[3]

[1] Department of Electronics and Informatics, Ryukoku University
Otsu, Shiga 520-2194, Japan
[2] NTT Communication Science Laboratories, NTT Corporation
Seika-cho, Kyoto 619-0237, Japan
[3] NTT Network Innovation Laboratories, NTT Corporation
Musashino, Tokyo 180-8585, Japan

Abstract. We aim to develop a technique to detect search engine optimization (SEO) spam websites. Specifically, we propose four methods for extracting the SEO spam entries from a given trackback network in blogspace that are based on fundamental metrics on a network. Using real data of trackback networks in blogspace, we experimentally evaluate the performance of the proposed methods, and demonstrate that the method of ranking entries based on average degrees of nearest neighbors can be a very promising approach for extracting SEO spam entries.

1 Introduction

Search engine optimization (SEO) is the process of increasing the amount of visitors to a Web site by ranking high in the search results of a search engine[1]. However, there are often SEO spam websites that contain little or no relevant content and whose aim is solely to increase their position in the search engine rankings. Such spamming involves obtaining more exposure for a website than it really deserves for a given search term, leading to unsatisfactory search experiences. Hence, it is an important research issue to develop a technique to detect SEO spam websites. With current search engines, the hyperlink strucure of the World Wide Web is widely exploited; for example, the "HITS" algorithm [6] and the "PageRank" algorithm [2] are well known. Thus, we consider the problem of detecting SEO spam websites based on the structures of link networks.

By contrast, considerable attention has recently been devoted to investigating weblogs (or *blogs*) [7],[5]. Here, blogs are personal on-line diaries managed by easy-to-use software packages, and they have spread rapidly through the World Wide Web. Someone who keeps a blog is called a *blogger*, and a collection of blogs with their links is referred to as *blogspace*. A blog consists of entries that include text, images, hyperlinks, and *trackbacks*. Compared with ordinary websites, one of the most important features of blogs is the existence of trackbacks. Unlike a hyperlink, one blogger can construct a link from an entry j of another blogger to

[1] see, `http://www.webopedia.com/TERM/S/SEO.html`

R. Khosla et al. (Eds.): KES 2005, LNAI 3684, pp. 723–729, 2005.

his entry i by creating a trackback on entry j. Thus, one can more easily create SEO spam entries in trackback networks. In this paper, we explore a method for extracting SEO spam entries from a trackback network.

Several investigations have been undertaken to identify the communities in a network by using graph-theoretic methods [3], [4]. Here, a community is defined as a collection of nodes in which each member node has more links to nodes within the community than to nodes outside the community. However, a set of SEO spam entries does not necessarily construct a community in a trackback network, since SEO spammers create trackbacks to their entries on normal entries that have many trackbacks in order to raise their rankings on search engines. This implies that the straightforward application of methods developed to identify communities are inadequate for our problem. To detect the SEO spam entries in a given trackback network, we propose four methods based on metrics on a network introduced in recent studies of complex network theory [9], [1], [8], and experimentally evaluate the performance of the proposed methods using real blog data.

2 Fundamental Metrics on Networks

We employ fundamental metrics on a network introduced in recent studies of complex network theory. We ignore the trackback direction for simplicity and treat a trackback network as an undirected graph. Therefore, throughout this work a network means an undirected graph.

2.1 Degrees and Average Degrees of Nearest Neighbors

The *degree* k_i of a node i in a network is defined as the number of links attached to node i [1]. One naive strategy for raising the rankings of blog entries on search engines is to create many trackbacks to those blog entries. Thus, we can naively consider that SEO spam entries should have high degrees.

By contrast, we can also consider that the blog entries with which an SEO spam entry connects should have high degrees. Thus, as studied in [8], we investigate the average degree $\bar{k}_i$ among the nearest neighbors of an entry i in a trackback network. We call $\bar{k}_i$ the *average NN degree* of entry i. Let $\mathcal{N}_i$ be the neighborhood of a node i in a network, that is, the set of nodes that have links to node i. Then, $\bar{k}_i$ is defined by

$$\bar{k}_i = \frac{1}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} k_j \ ,$$

where $|\mathcal{N}_i|$ denotes the number of elements in the set $\mathcal{N}_i$.

2.2 Clustering Coefficients

The *clustering coefficient* C_i of a node i in a network is defined by

$$C_i = \frac{2b_i}{k_i(k_i - 1)} \ ,$$

where b_i is the number of direct links connecting the nodes in the neighborhood $\mathcal{N}_i$ of node i [9]. Note that C_i reflects the probability that two friends of node i are friends themselves. We can naively consider that SEO spam entries should have high clustering coefficients in a trackback network.

3 Proposed Methods

We consider extracting the SEO spam entries from a given trackback network by ranking the entries in the network according to the level of SEO.

We propose the following four ranking methods based on the metrics introduced in the previous section. Let $r_1(i), r_2(i), r_3(i)$ and $r_4(i)$ be the evaluation functions of Methods 1, 2, 3 and 4, respectively, for measuring the SEO level of each node i. Then these functions are defined by $r_1(i) = C_i$, $r_2(i) = k_i$, $r_3(i) = C_i \log k_i$ and $r_4(i) = \bar{k}_i$. Figure 1 shows the kind of node in a network that is regarded as having a high SEO level for each method, i.e., a node with a high clustering coefficient for Method 1 (see, Fig. 1 (1)); a node with a high degree for Method 2 (see, Fig. 1 (2)); a node that has both a high clustering coefficient and a high degree for Method 3 (see, Fig. 1 (3)); and a node such that its nearest neighbors have high degrees for Method 4 (see, Fig. 1 (4)). Note that since the magnitude of the degree is generally much larger than that of the clustering coefficient, we performed a logarithmic transformation on the degree for Method 3.

4 Experimental Evaluation

4.1 Data Acquisition

Although there are a large number of blog entries in blogspace, many of them have no trackbacks. Namely, it is hard to obtain random samples of large connected trackback networks.

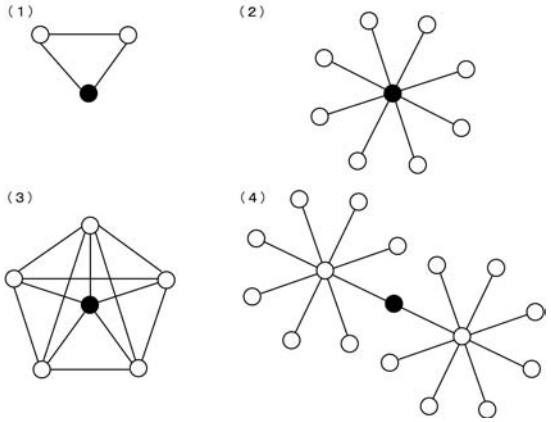

Fig. 1. Examples of nodes with high SEO levels for the proposed methods. The filled circles show examples of nodes with a high SEO level. (1)-(4) show examples for Methods 1-4, respectively

Then, we exploited the blog "Theme salon of blogs[2]", where blog users can recruit trackbacks of other bloggers by registering interesting themes. By tracing ten steps ahead the trackbacks from the blog entries for a theme in the "Theme salon of blogs", we collected a large connected trackback network. Note that the entries in the network were not restricted to the theme first chosen due to frequent topic drifts, and thus had a variety of topics. Namely, we might consider that this collection procedure could produce a reasonably random sampling of a large connected trackback network from blogspace.

4.2 Definition of SEO Spam Entries

We treated blog entries that had been participants in certain well-known SEO contests in Japan as SEO spam entries. In these SEO contests, SEO devotees compete for search engine rankings in a search for a specified keyword such as "Goggole" or "Deskedgar", where these words are artifacts for the contests. We defined an entry as an SEO spam entry if it has the banner (link farm) "Trackback OK, Goggole", the banner "Trackback OK, Deskedgar", or one of the following keywords in the blogger name, entry name, or description section: "Goggole", "Deskedgar", "Nama sanargi", "Yahhyoi", "Ponesonic", and "Dennou Purion".

4.3 Performance Measures

We quantified the perfomance of the proposed methods in terms of *F-measure* and *precision*, which are widely used in information retrieval.

Let S denote the set of SEO spam entries in a trachback network. We fix a method for extracting the SEO spam entries from the network. For any positive integer r, let M_r denote the set of the top r entries extracted by the method. Then, the F-measure $F(r)$ and the precision $P(r)$ of the method for ranking r are defined by

$$F(r) = \frac{2|M_r \cap S|}{|M_r| + |S|}, \quad P(r) = \frac{|M_r \cap S|}{|M_r|}.$$

Note that $F(r)$ quantifies how close the sets M_r and S are. Note also that the higher the value $P(r)$ is, the lower the detection error is.

4.4 Performance Evaluation

We describe our experimental results using data collected from the theme "Introduction of Special Sites" in the "Theme salon of blogs". Similar results were obtained by using data collected from other themes like "News for Smiling".

Then, the total numbers of blog entries and trackbacks were $9,338$ and $187,128$, respectively. By our definition described above, the number of SEO spam entries was $1,395$. Table 1 shows the fundamental statistics related to

[2] http://blog.goo.ne.jp/usertheme/

the proposed methods. Namely, the means of C_i, k_i and $\bar{k}_i$ are respectively displayed for the set of SEO spam entries and the others. Table 1 implies that the clustering coefficient, degree, and average NN degree of an SEO spam entry are generally larger than those of a non-SEO spam entry. Namely, these results justify applying the proposed methods.

Table 1. Measurement of the means of C_i, k_i and $\bar{k}_i$ for the set of SEO spam entries and the set of non-SEO spam entries

	$\langle C_i \rangle$	$\langle k_i \rangle$	$\langle \bar{k}_i \rangle$
SEO spam entries	0.50317	63.833	176.97
Non-SEO spam entries	0.27830	6.4297	24.267

Figures 2 and 3 respectively display F-measure $F(r)$ and precision $P(r)$ with respect to ranking r for the proposed methods. Here, when plural entries have the same score, the F-measure $F(r)$ and the precision $P(r)$ are not plotted until all such entries are included in the set M_r of ranking r. For example, both the F-measure and the precision graphs begin at $r = 828$ in Method 1, since there were 828 top entries.

Figure 2 shows that Method 4 provided the highest level of perfomance followed by Methods 2 and 3. Method 1 was the worst. In particular, the F-measure of Method 4 was extremely high with a value of over 80% around $r = 1,395$ (the

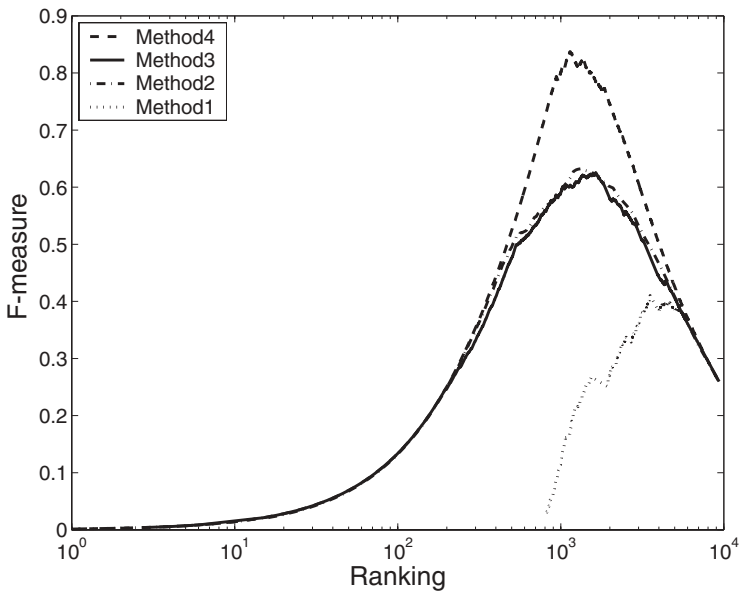

Fig. 2. Peformance evaluation of the proposed methods on F-measure. The dotted, dash-dotted, solid and dashed lines indicate the results for Methods 1-4,respectively

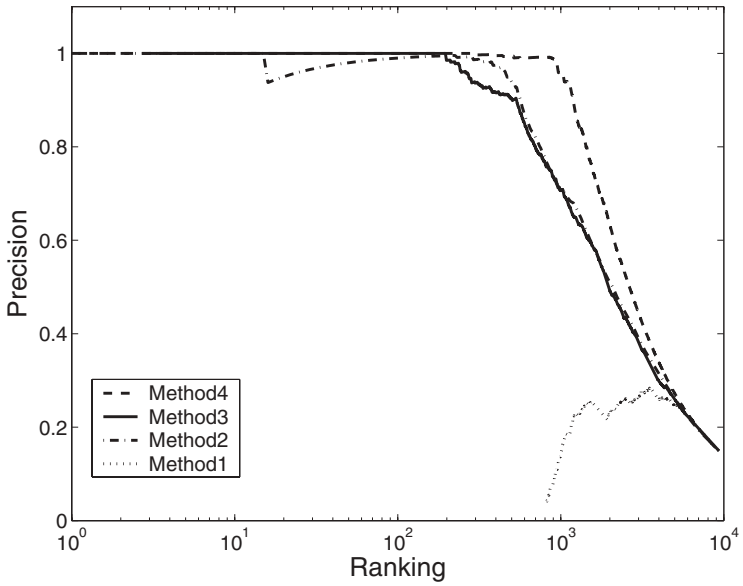

Fig. 3. Evaluation of the precision of the proposed methods. The dotted, dash-dotted, solid and dashed lines indicate the results for Methods 1-4,respectively

number of SEO spam entries). Moreover, Figure 3 shows that Method 4 was extremely precise. In particular, the value was 100% at $r = 289$ and over 90% around $r = 1,000$. These results imply that the method of ranking entries based on average NN degrees can be a very promising approach for extracting the SEO spam entries from a trackback network. We consider that this is because as a discriminative characteristic, SEO spam entries are likely to be linked to many entries with high degrees. Incidentally, Method 3 extracted "adult blog entries" as well as SEO spam entries. This suggests that adult blog entries are likely to produce relatively dense connections among them regardless of SEO spamming. Thus, Figs.2 and 3 indicate that the method of ranking entries based on both clustering coefficients and degrees is a promising approach for extracting general spam entries.

5 Conclusion

We proposed four methods based on fundamental metrics for extracting the SEO spam entries from a given trackback network and evaluated their performance experimentally. Using a connected trackback network collected by tracing ten steps ahead of the trackbacks from the blog entries for a theme in the "Theme salon of blogs", we experimentally demonstrated that the method of ranking entries based on average NN degrees can be a very promising approach to extract SEO spam entries. Moreover we showed that the method of ranking entries based on both clustering coefficients and degrees can be a promising way to extract general spam entries.

By contrast, the next important task is to undertake an extensive verification of our methods with various real blog data. To this end, we will need more sophisticated data collection processes. However, we have already made substantial progress, and we are encouraged by our initial results.

Acknowledgements

This work was partly supported by NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation. We express our sincere graditude to Naonori Ueda for his encouragement and useful suggestions. We thank Hirofumi Fujimoto for performing the experiments.

References

1. Barabási, A.-L. and Albert, R., Emergence of scaling in random networks, *Science*, **286** (1999) 509–512.
2. Brin, S. and Page, L., The anatomy of a large scale hypertextual Web search engine, In *Proceedings of the Seventh International World Wide Web Conference* (1998) 107–117.
3. Flake, G.W., Lawrence, S. and Giles, C.L., Efficient identification of Web communities, In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2000) 150–160.
4. Girvan, M. and Newman, E.J., Community structure in social and biological networks, *Proceedings of the National Academy of Sciences of the United States of America*, **99** (2002) 7821–7826.
5. Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A., Information diffusion through blogspace, In *Proceedings of the 13th International World Wide Web Conference* (2004) 491–501.
6. Kleinberg, J., Authoritative sources in a hyperlinked environment, In *Proceedings of the Ninth ACM-SIAM Symposium on Discrete Algorithms* (1998) 668–677.
7. Kumar, R., Novak, J., Raghavan, P., and Tomkins, A., On the bursty evolution of Blogspace, In *Proceedings of the 12th International World Wide Web Conference* (2003) 568–576.
8. Pastor-Satorras, R., Vázquez, A., and Vespignani, A., Dynamical and correlation properties of the Internet, *Physical Review Letters*, **87** (2001) 258701.
9. Watts, D.J. and Strogatz, S.H., Collective dynamics of 'small-world' networks, *Nature*, **393** (1998) 440–442.