Collecting and Analyzing Japanese Splogs based on Characteristics of Keywords

Yuuki Sato[⋆] **Takehito Utsuro**[⋆] **Tomohiro Fukuhara**[‡]

Yasuhide Kawada[◇] **Yoshiaki Murakami**[◇] **Hiroshi Nakagawa**[‡] **Noriko Kando**[♯]

[⋆]University of Tsukuba, Tsukuba, 305-8573, JAPAN [‡]University of Tokyo, Kashiwa/Tokyo, 277-8568/113-0033, JAPAN

[◇]Navix Co., Ltd., Tokyo, 141-0031, JAPAN [♯]National Institute of Informatics, Tokyo, 101-8430, JAPAN

Abstract

This paper focuses on analyzing (Japanese) splogs based on various characteristics of keywords contained in them. We estimate the behavior of spammers when creating splogs from other sources by analyzing the characteristics of keywords contained in splogs. Since splogs often cause noises in word occurrence statistics in the blogosphere, we assume that we can efficiently collect splogs by sampling blog homepages containing keywords of a certain type on the date with its most frequent occurrence. We manually examine various features of collected blog homepages regarding whether their text content is excerpt from other sources or not, as well as whether they display affiliate advertisement or out-going links to affiliated sites. Among various informative results, it is important to note that more than half of the collected splogs are created by a very small number of spammers.

Introduction

Spam blogs or splogs are blogs hosting spam posts, created using machine generated or hijacked content for the sole purpose of hosting advertisements or raising the PageRank of target sites. (Kolari, Joshi, & Finin 2006) reported that for English blogs, around 88% of all pinging URLs (i.e., blog homepages) are splogs, which account for about 75% of all pings. Based on this estimation, splogs can cause problems including the degradation of information retrieval quality and the significant waste of network and storage resources. Several previous works (e.g., (Kolari, Joshi, & Finin 2006)) reported important characteristics of splogs such as ping time series, in-degree/out-degree distributions, and typical words in splogs. In the context of semi-automatically collecting web spams including splogs, (Wang *et al*. 2007) discuss how to collect spammer-targeted keywords to be used when collecting a large number of web spams efficiently.

Unlike those previous works, this paper focuses on analyzing (Japanese) splogs based on various characteristics of keywords contained in them. As has been often noted in the previous works, text content of splogs is mostly excerpted from other sources such as news articles, blog articles (posts), advertisement pages, and other web texts. Considering this fact, in this work, we estimate the behavior of spammers when creating splogs from other sources by analyzing the characteristics of keywords contained in splogs.

The characteristics of keywords to which we pay attention in this paper is the importance of their information value as well as the duration of the keywords' having the information value. Figure 1 shows the keyword map we use for characterizing keywords, as well as 22 keywords that have balanced positions on this map (placed totally by intuition), and are used for collecting splogs. The vertical axis of the map denotes the importance of the information value that each keyword has, while its horizontal axis denotes the duration of each keyword's having the information value. Keywords with high information value are typically reported in news as social/political/economical issues, while those with low information value are typically issues regarding entertainment or celebrity, or high paying adsense keywords. On the other hand, keywords with short term duration include seasonal ones and those related to temporary events, while those with long term duration include organization names with a long history such as political parties and country names, or those related to permanent issues such as health and beauty.

Next, since splogs often cause noises in word occurrence statistics in the blogosphere, we collect splogs by sampling blog homepages[1] containing keywords of a certain type on the date with its most frequent occurrence[2] during the year 2007, and then by manually judging whether each blog homepage to be a splog or an authentic blog[3]. We then manually examine various features of collected blog homepages regarding whether their text contents are excerpts from other sources or not, as well as whether they display affiliate advertisement or out-going links to affiliated sites. Among various informative results of our analysis, it is important to note that more than half of the collected splogs are created by a very small number of spammers, and hence, the analysis reported in this paper is strongly affected by the choices

[1]For collecting the Japanese blog data, we use the system called KANSHIN (Fukuhara, Utsuro, & Nakagawa 2007) which collects blog articles (posts) written in Chinese, Japanese, Korean, and English, and has 3.6 million blog homepages and 193 million articles registered for Japanese since March 18th, 2004.

[2]Here, we prefer blog homepages with more posts per day than those with fewer posts per day.

[3]A blog homepage is a *splog* if it does not contain originally written text, or even if it does, having links to affiliated sites.

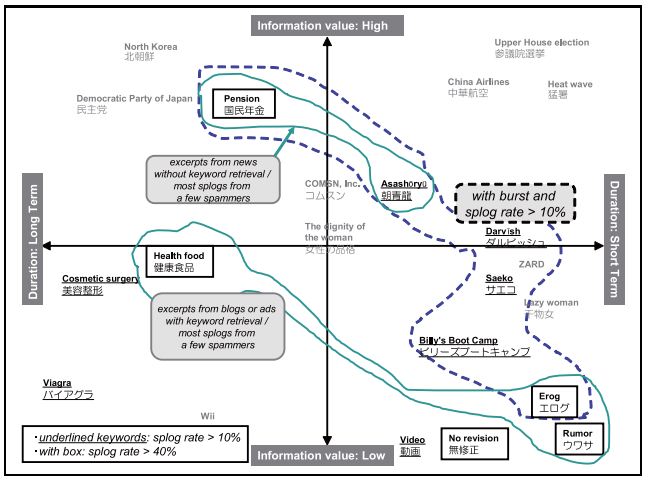

Figure 1: Keyword Map with Splog Analysis Results

of those spammers when they create those splogs.

Preliminary Results of Analyzing Splogs

Blog Hosts Statistics

As shown in Figure 2, in our Japanese blog homepage data set, more than 85% of splogs are from the top three hosts. It is estimated that those hosts with high splog rates pay less cost of manually removing splogs than those with low splog rates. As we show in the analysis of the next section, it is observed that a very small number of spammers actually create substantial number of splog homepages on those three hosts, and this increases the splog rates of those hosts.

Relations between Characteristics of Keywords and Splogs

Based on feature analysis of splogs in the entire splog data set, we examine correlation of those splog features and characteristics of keywords with splog rates higher than 10%. Major conclusions of this analysis can be summarized as below, some of which are also noted in Figure 1.

(1) The most important fact to note here is that, for four out of the five keywords with splog rate over 40%, most splog homepages are created by a very small number of spammers. Splogs containing these four keywords actually amount to more than half of the entire splog data set. This fact is very important because the following analysis is strongly affected by the choices of those spammers when they create those splogs.

(2) As shown in the map in Figure 1, most of the keywords placed in the upper half of the map have low splog rates. This means that splogs tend to contain keywords with low information value more often than those with high information value. *kokumin-nenkin(national pension)* and *Asashō-ryū* are with exceptionally high splog rates, though this statistics is strongly affected by the choices of less than five spammers. Those spammers posted splog posts on certain dates, where the splog articles are created from the excerpts of the news reports on those dates. Those excerpts occasionally include scandal reports closely related to the two keywords.

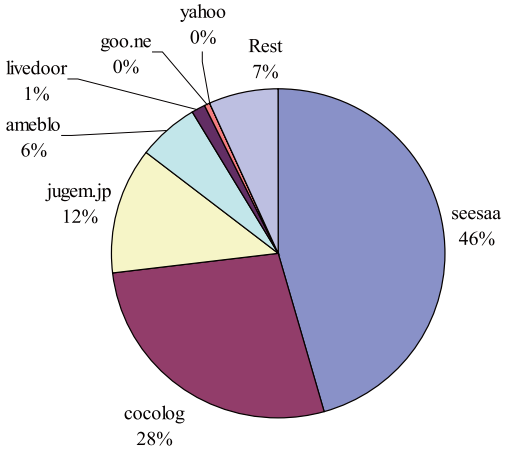

Figure 2: Blog Host Distribution in the Splog Homepage Data Set

(3) The three keywords *uwasa (rumor)*, *erogu* (*Erog*, adult content blog), and *kenko-shokuhin (health food)* correspond to another group of splogs created by a very small number of spammers. In the case of these keywords, the spammers posted splog posts, where the splog articles are created from the excerpt of other blogs and advertisements, but not news articles, by retrieving them with certain keywords.

(4) Among those six keywords with burst and their splog rates over 10%, *erogu* (*Erog*, adult content blog) is exceptional since its burst seems to happen simply because blog hosts started manually removing splogs including this keyword from certain period during our observation.

Conclusion

This paper focused on analyzing (Japanese) splogs based on various characteristics of keywords contained in them. It is important to note that more than half of the collected splogs are created by a very small number of spammers. Future works include further analysis of splogs by integrating with other features studied in the previous works (Kolari, Joshi, & Finin 2006), such as characteristic words in splogs, in-degree/out-degree distributions, and ping time series. Next, we plan to apply existing splog detection techniques (Kolari, Finin, & Joshi 2006) to our splog data set, and then to develop a splog detector with high accuracy.

References

Fukuhara, T.; Utsuro, T.; and Nakagawa, H. 2007. Cross-lingual concern analysis from multilingual weblog articles. In *Proc. 6th Inter. Workshop on Social Intelligence Design*, 55–64.

Kolari, P.; Finin, T.; and Joshi, A. 2006. SVMs for the Blogosphere: Blog identification and Splog detection. In *Proc. 2006 AAAI Spring Symp. Computational Approaches to Analyzing Weblogs*, 92–99.

Kolari, P.; Joshi, A.; and Finin, T. 2006. Characterizing the splogosphere. In *Proc. 3rd Ann. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Wang, Y.; Ma, M.; Niu, Y.; and Chen, H. 2007. Spam double-funnel: Connecting web spammers with advertisers,. In *Proc. 16th WWW Conf.*, 291–300.