Analysing Features of Japanese Splogs and Characteristics of Keywords

Yuuki Sato Takehito Utsuro University of Tsukuba, Tsukuba, 305-8573, JAPAN

> Yoshiaki Murakami Navix Co., Ltd., Tokyo, 141-0031, JAPAN

Tomohiro Fukuhara University of Tokyo, Kashiwa 277-8568, JAPAN

Hiroshi Nakagawa University of Tokyo, Tokyo, 113-0033, JAPAN Yasuhide Kawada Navix Co., Ltd., Tokyo, 141-0031, JAPAN

Noriko Kando 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 (manually) 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.

Categories and Subject Descriptors

H.3.0 [INFORMATION STORAGE AND RETRIEVAL]: General

General Terms

Reliability

Keywords

Blog analysis, splog, time series characteristics of keywords, keyword bursts

1. INTRODUCTION

Weblogs or blogs are considered to be one of personal journals, market or product commentaries. While traditional search engines continue to discover and index blogs, the blogosphere has produced custom blog search and analysis en-

Copyright 2008 ACM 978-1-60558-159-0 ...\$5.00.

gines, systems that employ specialized information retrieval techniques. There are several previous works and services on blog analysis systems. [13] proposed a system called *blog-Watcher* that collects and analyzes Japanese blog articles. [6] proposed a system called *BlogPulse* that analyzes trends of blog articles. With respect to blog analysis services on the Internet, there are several commercial and non-commercial services such as *Technorati*¹, *BlogPulse*², *kizasi.jp*³, and *blog-Watcher*⁴. With respect to multilingual blog services, *Globe of Blogs*⁵ provides a retrieval function of blog articles across languages. *Best Blogs in Asia Directory*⁶ also provides a retrieval function for Asian language blogs. *Blogwise*⁷ also analyzes multilingual blog articles.

As with most Internet-enabled applications, the ease of content creation and distribution makes the blogosphere spam prone [7, 1, 10, 12, 9]. 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. [10] 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, as stated in [1, 11], splogs can cause problems including the degradation of information retrieval quality and the significant waste of network and storage resources. Several previous works [10, 12, 9] reported important characteristics of splogs. [12] reported characteristics of ping time series, in-degree/out-degree distributions, and typical words in splogs found in TREC⁸ Blog06 data collection. [10, 9] also reported the results of analyzing splogs in the BlogPulse data set. In the context of semi-automatically collecting web spams including splogs, [16] 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 [14]. As has been often noted in the previous works, text content of splogs is mostly ex-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

AIRWeb '08, April 22, 2008 Beijing, China.

¹http://technorati.com/

²http://www.blogpulse.com/

³http://kizasi.jp/ (in Japanese)

⁴http://blogwatcher.pi.titech.ac.jp/ (in Japanese)

⁵http://www.globeofblogs.com/

⁶http://www.misohoni.com/bba/

⁷http://www.blogwise.com/

⁸http://trec.nist.gov/

1	able 1: Features for v	Characterizing Spiogs and their Rates in Spiog Data Set	
Feature Types	Features	Descriptions	Rate in Splogs (%)
	links to affiliated	Blog articles (posts) contain sufficiently many out-going links to	80.5
	sites	affiliated sites except for the out-going links that the blog hosts	00.0
51065		automatically add to individual blog homenages and blog posts	
Affiliate	advertisement arti-	Blog articles (posts) themselves contain sufficiently many ad-	31.0
1 tilliate	clos (posts)	vortisements except for the advertisements that the blog bests	01.0
	cies (posts)	automatically add to individual blog homopages and blog nosts	
Fosturos	articles (posts) with	Blog articles (posts) contain adult content	81
reatures	adult content	blog articles (posts) contain aduit content.	0.1
	keywords with popup	Certain blog hosts have facilities of automatically adding popup	42.1
	advertisement	advertisements to keywords.	
	excerpt from news ar-	Text content is automatically or manually excerpted from news	14.3
	ticles	articles.	
Content	excerpt from blog ar-	Text content is automatically or manually excerpted from other	70.8
	ticles (posts) or other	blog articles (posts), or web texts other than news articles and	
	web texts	advertisement pages.	
Source	excerpt from adver-	Text content is automatically or manually excerpted from cer-	27.1
	tisement pages	tain advertisement pages.	
Features	originally written	Spammers write original splog texts.	2.9
	texts		
	meaningless sequence	Most of them are so called word salad spam text [2] and are	3.6
	of words	automatically generated.	
	excerpt from other	Text content is automatically or manually excerpted from other	12.7
	sources, selected	sources without keyword retrieval. Typical cases are excerpt	
	without keyword	from news articles or blog posts on the same date or close dates.	
	retrieval		
Creation	excerpt from other	Text content is automatically or manually retrieved from other	49.5
	sources, retrieved	sources with a keyword varying day by day, and then excerpted.	
	with a keyword		
	varying day by day		
Procedure	excerpt from other	For a blog homepage, all of its text content is excerpt, which are	36.9
	sources, retrieved	automatically or manually retrieved from other sources with a	
	with a single key-	single keyword throughout all of its posts.	
	word throughout a		
	blog homepage		
Features	keyword stuffed	Blog articles (posts) contain lists of keywords for SEO purposes.	11.5
	blog [9]		
	automatically gener-	Most of them are so called word salad spam text [2], which is	4.5
	ated text	a mixture of seemingly meaningful words that together signify	
		nothing. Sometimes, connecting several sentences each of which	
		is excerpted from other source.	

cerpted 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 a keyword to which we pay attention in this paper is whether the keyword is of public/private concern as well as the duration of people's concern to the keyword. Furthermore, 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 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 of those spammers when they create those splogs.

2. PROCEDURE OF CREATING SPLOGS

Text content of splogs is mostly excerpted from other sources such as news articles, blog articles (posts), advertisement pages, and other web texts. In any case, splogs have commercial intention — they display affiliate advertisement or out-going links to affiliated sites. For this purpose, splogs are usually created by searching for up-to-date content from other sources and by excerpting them. This procedure of creating splogs can be roughly divided into the following two cases:



Figure 1: Time Series Characteristics of Keyword Occurrence Statistics in Splogs / Authentic Blogs

- i) excerpting text content from news articles or blog posts on the same date or close dates *without keyword retrieval*,
- ii) excerpting text content by *retrieving* them from other sources with certain keywords.

Splog posts created by the first procedure just a few days before the current date tend to contain up-to-date text content which are originally from quite recent news articles or blog posts. On the other hand, for splogs created by the second procedure, spammers usually carefully choose keywords for retrieving text content from other sources such as news articles and blog posts. They tend to choose high paying adsense⁹ keywords.

3. FEATURES FOR CHARACTERIZING SPLOGS

This section describes the features for characterizing Japanese splog homepages manually collected by the procedure of section 5.3.

As we summarize in Table 1, this paper considers the following three types of features for splogs, namely, 1) affiliate features, 2) content source features, and 3) creation procedure features. For each of these three feature types, Table 1 lists several binary features each of which denotes whether the given splog homepage has the designated characteristics or not. Here, note that features of the same type are independent of each other and hence are not necessarily disjoint. Also note that most of those features are for the use in manual examination of splogs, and hence, it is not necessarily meant to automatically detect them.

3.1 Affiliate Features

Among the three feature types, first we describe *affiliate features*. As introduced in [10, 9], splogs are generated with two often overlapping motives, namely, creation of fake blogs for the purpose of hosting profitable advertisement, and unjustifiably increasing the ranking of affiliated sites. Since both motives are deeply related to affiliate advertising, in this paper, we consider features of splogs regarding issues of affiliates. As the affiliate features, we manually examine the following four points:

- i) whether the blog article (posts) contain out-going links to affiliated sites,
- ii) whether the blog article (posts) themselves contain advertisements,
- iii) whether blog articles (posts) contain adult content¹⁰,
- iv) whether blog articles (posts) contain popup advertisements automatically added to certain keywords.

3.2 Content Source Features

Second, one of the important characteristics of splogs is that their text content is mostly excerpted from other sources such as news articles, blog articles (posts), advertisement pages, and other web texts. In order to estimate the mechanism of creating splogs, we manually examine the content source of splogs and classify them according to the following five features, namely, *content source features*:

- i) excerpt from news articles,
- ii) excerpt from blog articles (posts) or other web texts,
- iii) excerpt from advertisement pages,
- iv) originally written texts,
- v) meaningless sequence of words such as word salad spam texts [2].

3.3 Creation Procedure Features

Furthermore, we estimate the procedures of searching the web for those excerpt and manually classify them according to the following five features, namely, *creation procedure features*:

- i) excerpt from other sources, selected without keyword retrieval, where typical cases are excerpt from news articles or blog posts on the same date or close dates,
- ii) excerpt from other sources, retrieved with a keyword varying day by day,
- iii) excerpt from other sources, retrieved with a single keyword throughout a blog homepage,
- iv) keyword stuffed blog [9],

⁹http://google.com/adsense

¹⁰Adult content is among the major target genres for affiliate advertising, while other major target genres include health food and slimming products, cosmetics, and finance. We regard blogs which contain adult content as more harmful than others, and record them with an independent feature.



Figure 2: A Keyword Map for Characterizing Keywords

 v) automatically generated text including word salad spam texts [2].

As the creation procedure features, we distinguish two major procedures of creating splogs, i.e.,

- a) excerpt from news articles or blog posts on the same date or close dates *without keyword retrieval*, and
- b) and excerpt by *retrieving* texts from other sources *with certain keywords*.

The former type corresponds to the feature i) above, while the latter to the features ii) and iii) above.

4. CHARACTERISTICS OF SPLOGS AND KEYWORDS

4.1 Time Series Characteristics of Keywords

Among the problems caused by splogs, this section discusses issues on noises in word occurrence statistics in the blogosphere. Figure 1 illustrates two typical cases of noises in time series keyword occurrence statistics, where (a) is the case of a keyword with burst, and (b) is the case of a keyword without burst. For both cases, keyword occurrences are mixture of those from authentic blogs and splogs. Without detecting and removing splogs, it is difficult to estimate real keyword occurrence statistics only in authentic blogs. For the case of the keywords with burst, especially, it is estimated that burst in splogs may be delayed from that in authentic blogs, because text content of splogs is mostly excerpt from other sources such as news articles and blog posts.

4.2 Keyword Map for Characterizing Keywords

This section introduces the keyword map of Figure 2 for characterizing keywords. The vertical axis of the map denotes whether each keyword is of public/private concern, while its horizontal axis denotes the duration of people's concern to each keyword. Keywords with public concern are typically reported in news as social/political/economical issues, while those with private concern 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.

On the map of Figure 2, 50 keywords that are balanced in their distribution on the map are placed, where the position of each keyword is determined totally by intuition. Those keywords vary in their time series characteristics of occurrence statistics, where some of them are with burst while others are not. Each of those keywords is intended to be used for retrieving blog (authentic blog and splog) homepages in the procedure of section 5.3. The major purpose of placing such various keywords onto a map like this is to simply examine the correlation between the characteristics of keywords and the rate of splogs among the blogs containing each keyword.

Table 2: Summary of Japanese Blog Data (at December 3rd, 2007, 0:00)

Γ	# of blog			current $\#$ of
	homepages	# of articles	# of days	articles per day
Γ	3,591,306	192,699,276	1,355	196,975

5. ANALYZING SPLOGS BASED ON CHAR-ACTERISTICS OF KEYWORDS

5.1 Motivations

This paper reports the results of analyzing the following three points after collecting blogs and then manually detecting splogs among them.

- 1. Features of splogs are manually examined according to those introduced in section 3.
- 2. According to the keyword map for characterizing keywords, various characteristics of keywords are manually examined, which include time series characteristics such as whether *with/without* burst.
- 3. Based on the results of examining above two points, we further analyze various correlation between characteristics of splogs and keywords. This analysis mainly includes the followings:
 - (a) correlation between the characteristics of keywords and the rate of splogs among the blogs containing each keyword. This will reveal the preference of spammers when choosing keywords.
 - (b) correlation between the characteristics of keywords and the splog creation procedures.

5.2 Japanese Blog Data

For collecting the Japanese blog data, we use the system called KANSHIN [3, 4, 5] which collects blog articles (posts) written in Chinese, Japanese, Korean, and English. The system has lists of blog homepages for each language. By using these lists, the system collects RSS^{11} and Atom feed files provided by blog homepages, and extracts keywords from feed files by using morphological analysis tools, and store keywords and articles in each database. The system uses several linguistic tools for extracting and indexing keywords from blog articles for each language. For Japanese, it uses a morphological analysis tool called Juman¹². The system provides users with functions for retrieving and analyzing articles.

Table 2 shows the summary of Japanese blog data stored in the system (checked at December 3rd, 2007). 3.6 million blog homepages and 193 million articles are registered for Japanese since March 18th, 2004.

5.3 Procedure of the Analysis

This section gives the specific procedure of collecting and analyzing splogs based on characteristics of keywords. The rough strategy of collecting splogs here is

to simply collect *blog homepages*, (i.e., *not* blog posts) which contain a given keyword and then,



Figure 3: Blog Host Distribution in the Splog Homepage Data Set

considering the features of splogs defined in section 3, to manually judge whether each of the collected blog homepages is a splog or an authentic blog.

Considering the result of a preliminary examination, we assume that, for keywords *with* burst, the rate of splogs among the blog homepages that contain those keywords may be higher on the burst date than on other dates. We further assume that, even for keywords *without* burst, the rate of splogs may be higher on the date with the most frequent occurrence in the blogosphere than other dates. Based on this observation, in order to collect sufficient number of splogs, for each keyword, we collect blog homepages containing the keyword on the date with its most frequent occurrence. Furthermore, also considering the result of a preliminary examination, we prefer blog homepages with more posts per day than those with fewer posts per day.

The following list summarizes the above procedure.

- 1. For each of the 50 keywords in Figure 2, we collect blog homepage URLs which contain the keyword on the date with its most frequent occurrence during the year 2007.
- 2. Among the collected URLs, we select the topmost 50 with respect to the number of posts per day. We further randomly select 60 URLs from the rest. This amount to 110 URLs in total, where the topmost 50 URLs are usually with more than three posts per day, while the remaining 60 URLs are with one or two posts per day.
- 3. For each of the collected URLs, an annotator judges whether each binary feature defined in section 3 holds or not.
- 4. Based on the above judgement, each URL is judged to be a splog or an authentic blog according to the following rule.
 - (a) If one of the followings holds for the given URL, then it is mostly¹³ splog.

¹¹Several references such as RDF Site Summary or Really Simple Syndication or Rich Site Summary exist.

¹²http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman. html

 $^{^{13}\}mathrm{By}$ "mostly", we mean that it is usually necessary to judge by considering the contents of each blog.

Table 3: Splog Rate per Blog Host

Blog Host		seesaa	cocolog	jugem.jp	ameblo	livedoor	goo.ne	yahoo	Rest	Total
# of Blog	Splog	192	142	54	24	3	21	0	26	442
Homepages	Authentic Blog	203	115	169	355	128	130	207	396	1703
	Total	395	257	223	379	131	131	207	422	2145
Splog Rate (%)		48.6	55.3	24.2	6.3	2.3	0.8	0.0	6.2	20.6

Table 4: Splog Rate, Professional Spammer Rate (from professional spammer / splog), # of Professional Spammers, and, Amateur Only Splog Rate (from amateur spammer / (from amateur spammer + non-splog)) (in descending order of splog rates, boldfaced: "splog rate > 10%, professional spammer rate > 50%", underlined: "amateur only splog rate ~ 20% or more, mostly with private concern")

Keyword	Splog Rate (%)	Professional Spammer Rate (%)	# of Professional Spammers	Amateur Only Splog Rate (%)
erog, adult content blog	89.2	92.4	3	38.5
rumor	88.1	94.8	1	27.8
national pension	58.1	90.2	2	12.0
<u>no revision</u>	40.9	18.5	1	36.1
health food	37.4	58.7	2	19.8
cosmetic surgery	24.4	14.3	2	21.7
Viagra	22.5	11.1	1	20.5
Darvish, a Japanese baseball player	22.1	0.0	0	22.1
video	19.1	0.0	0	19.1
Asashō-ryū, a sumo wrestler	15.2	80.0	2	3.4
Billy's Boot Camp	15.1	0.0	0	15.1
Saeko, a Japanese actress and Darvish's wife	14.3	14.3	1	12.2
COMSN, Inc., elderly care business company with a scandal	6.9	71.4	2	2.1
ZARD (a Japanese female singer, accidentally died)	4.7	20.0	1	3.8
China Airlines	4.7	20.0	1	3.8
North Korea	2.9	100.0	1	0.0
Wii (a video game console of Nintendo)	2.8	66.7	1	1.0
heat wave	2.8	33.3	1	1.9
"The dignity of the woman", the title of a book	2.0	0.0	0	2.0
a Japanese slang word for "lazy woman"	1.8	50.0	1	0.9
Upper House election	0.0	0.0	0	0.0
Democratic Party of Japan	0.0	0.0	0	0.0
Total	20.5	61.5	10	9.0

i. The feature "originally written text" does not hold.

- ii. The feature "originally written text" holds and at least one of the features "links to affiliated sites", "advertisement articles (posts)", or "articles (posts) with adult content" holds.
- (b) Otherwise, the given URL is an *authentic blog*.
- 5. Finally, we analyze the correlation between characteristics of keywords and the distribution of features manually annotated to splogs.

6. PRELIMINARY RESULTS OF ANALYZ-ING SPLOGS

This section discusses preliminary results of analyzing Japanese splogs based on characteristics of keywords, features of splogs, as well as other features which can be automatically analyzed such as blog hosts distribution. We further analyze the correlation between characteristics of keywords and the feature distribution of splogs. Here, note that the results shown below are preliminary in that they are for 22 keywords out of the 50 on the map of Figure 2.

6.1 Blog Hosts Statistics

As can be clearly seen from Figure 3, in our Japanese blog homepage data set, more than 88% of splogs are from the top three hosts. Furthermore, as shown in Table 3, for the top two hosts, about half of the blog homepages are $splogs^{14}$. 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 argue in 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.

6.2 Relations between Characteristics of Keywords and Splogs

¹⁴Due to errors in the procedure of collecting blog URLs for judging splog/authentic blog distinction, for the moment, we do not have 110 blogs URLs in total for several keywords.

	# of	Fea	tures of Splogs (in Tabl		
ID	Splogs	Affiliate	Content Source	Creation Procedure	Keywords
1	115 (42.5%)	links to affiliated sites, popup adver- tisement	blog or other web texts	retrieved with a sin- gle keyword	rumor, no revision, cosmetic surgery, Asashō-ryū, Saeko, China Airlines, COMSN, Inc., ZARD, heat wave, Wii, North Korea, "lazy woman"
2	56 (20.6%)	links to affiliated sites	blog or other web texts	retrieved with a key- word varying day by day	erog
3	$ \begin{array}{c} 30 \\ (11.0\%) \end{array} $	links to affiliated sites	news articles, adver- tisement pages	selected without key- word retrieval	national pension, COMSN, Inc.
4	26 (9.6%)	links to affiliated sites, advertisement articles, popup advertisement	blog or other web texts, advertisement pages	retrieved with a key- word varying day by day	national pension
5	20 (7.4%)	links to affiliated sites, advertisement articles	advertisement pages	retrieved with a key- word varying day by day, keyword stuffed blog	health food
6	$10 \\ (3.7\%)$	links to affiliated sites, adult content, popup advertisement	news articles, blog or other web texts	selected without key- word retrieval	erog, Asasho-ryu,
7~10	$ \begin{array}{c} 15 \\ (5.5\%) \end{array} $	<u> </u>		<u> </u>	erog, health food, Viagra, cos- metic surgery,
Total	272		—	—	

 Table 5: 10 Professional Spammers identified in our Splog Data Set

Next, for each of the 22 keywords, Table 4 gives splog rates in the blog homepages collected with the keyword, in descending order of splogs rates. In the table, those 22 keywords are divided into three groups, i.e., those with splog rates higher than 30%, those with splog rates $30 \sim 10\%$, and the rest. We further count occurrences of features of splogs in the entire splog data set, and list their rates in the splog data set as in the rightmost column of Table 1. Based on this feature analysis, we examine correlation of those splog features and characteristics of keywords with splog rates higher than 10%.

Furthermore, we judged whether two splogs are created by an identical spammer when their html layouts are similar¹⁵, and then grouped those splogs from an identical spammer. In this paper, we name those spammers each of whom created more than one splogs in our data set as *professional* spammers, while we also name those remaining spammers each of whom created only one splog in our data set as amateur spammers. With this judgement, we can identify 10 professional spammers in our splog data set (summarized in Table 5), where out of the total 442 splog homepages, 272 (61.5%) can be regarded as created by those 10 professional spammers. Based on this professional/amateur spammer analysis, for each keyword, Table 4 shows rate of splog homepages being created by one of the 10 professional spammers Table 4 also shows the number of professional spammers observed for each keyword, as well as splog rates after removing those created by professional spammers (amateur only splog

¹⁵Our next plan is to employ the technique presented in [15], so that we can automatically group splog homepages into the 10 groups shown here. rate).

Major conclusions of this analysis can be summarized as below, some of which are also noted in the map of the 22 keywords in Figure 4.

(1) The most important fact to note here is that, for four out of the five keywords with splog rate over 30%, most splog homepages are created by professional 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 professional spammers in creating those splogs.

(2) As can be seen from the map in Figure 4, 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 private concern more often than those with public concern. "National pension" and "Asashō-ryū" are with exceptionally high splog rates, though this statistics is strongly affected by the choices of professional spammers. Those spammers posted splog posts on certain dates, where the splog articles are created from the excerpts of the news reports and blog posts on those dates. Those excerpts occasionally include scandal reports closely related to the two keywords.

(3) The three keywords "rumor", "erog, adult content blog", and "health food", correspond to another group of splogs created by professional 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.

7. CONCLUSION



Figure 4: Keyword Map with Splog Analysis Results

This paper focused on analyzing (Japanese) splogs based on various characteristics of keywords contained in them. 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 professional spammers. Future works include further analysis of splogs by integrating with other features studied in the previous works [12, 10, 9], such as characteristic words in splogs, in-degree/outdegree distributions, and ping time series. Next, we plan to apply existing splog detection techniques [11, 8] to our splog data set, and then to develop a splog detector with high accuracy. Splogs/authentic blogs collected in this work are also useful for analyzing characteristics of keywords in a much larger scale, simply by automatically collecting a much larger number of keywords, and then measuring correlation between splogs and each keyword.

8. **REFERENCES**

- Wikipedia, Spam blog. http://en.wikipedia.org/wiki/ Spam_blog.
- [2] Wikipedia, Word salad (computer science). http://en. wikipedia.org/wiki/Word_salad_%28computer_science%29.
- [3] T. Fukuhara, T. Murayama, and T. Nishida. Analyzing concerns of people using Weblog articles and real world temporal data. In *Proceedings of WWW 2005 2nd Annual* Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2005.
- [4] T. Fukuhara, H. Nakagawa, and T. Nishida. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. In *Proceedings of ICWSM*, pages 271–272, 2007.
- [5] T. Fukuhara, T. Utsuro, and H. Nakagawa. Cross-lingual concern analysis from multilingual weblog articles. In A. Nijholt, O. Stock, and T. Nishida, editors, *Proceedings* of the 6th International Workshop on Social Intelligence Design, pages 55–64, 2007.

- [6] N. Glance, M. Hurst, and T. Tomokiyo. Blogpulse: Automated trend discovery for Weblogs. In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004.
- [7] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In Proc. 1st AIRWeb, pages 39–47, 2005.
- [8] P. Kolari, T. Finin, and A. Joshi. SVMs for the Blogosphere: Blog identification and Splog detection. In Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs, pages 92–99, 2006.
- [9] P. Kolari, T. Finin, and A. Joshi. Spam in blogs and social media. In *Tutorial at ICWSM*, 2007.
- [10] P. Kolari, A. Joshi, and T. Finin. Characterizing the splogosphere. In Proceedings of WWW 2006 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006.
- [11] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *Proc. 3rd AIRWeb*, pages 1–8, 2007.
- [12] C. Macdonald and I. Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow, Department of Computing Science, 2006.
- [13] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining Japanese weblogs. In WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, pages 320–321. ACM Press, 2004.
- [14] Y. Sato, T. Utsuro, T. Fukuhara, Y. Kawada, Y. Murakami, H. Nakagawa, and N. Kando. Collecting and analyzing Japanese splogs based on characteristics of keywords. In *Proc. ICWSM*, pages 218–219, 2008.
- [15] T. Urvoy, T. Lavergne, and P. Filoche. Tracking Web spam with hidden style similarity. In *Proc. 2nd AIRWeb*, pages 25–30, 2006.
- [16] Y. Wang, M. Ma, Y. Niu, and H. Chen. Spam double-funnel: Connecting web spammers with advertisers,. In Proc. 16th WWW Conf., pages 291–300, 2007.