Analyzing Trackback Usage as a Inspection of Weblog Data Quality for Weblog Mining

Shinsuke Nakajima[1], Katsumi Tanaka[2] and Shunsuke Uemura[3]

[1] Graduate School of Information Science, Nara Institute of Science and Technology.
8916-5 Takayama-cho Ikoma Nara 630-0101, Japan
shin@is.naist.jp
[2] Dept. of Social Informatics, Kyoto University.
Yoshida Honmachi Sakyo-ku Kyoto 606-8501, JAPAN,
ktanaka@i.kyoto-u.ac.jp
[3] Faculty of Informatics, Nara Sangyo University.
3-12-1 Tateno-Kita, Sango, Ikoma, Nara 636-8503, Japan,
UemuraShunsuke@nara-su.ac.jp

Abstract. Recently, blogs have become widely used as tools for putting out information quickly and easily. Thus, the Web has become not only a place for getting information, but also a place for communication. It can be said that blogs have changed the way people use the Internet and become the mirror of public opinion. Some researchers do blog analysis such as blog community analysis and reputation analysis using blog data. It is known that the link structure among blog entries considerably influences the formation of blog communities on blogspace. Thus, it is very important to investigate hyperlinks and trackback links in order to understand characteristics of blog communities and blogger behavior. However, most researchers do not focus on trackback links, despite their importance in understanding the relations between blog entries. Therefore, we analyze trackback usage in order to inspect weblog data quality for weblog mining and investigate their importance in understanding blogspace for blogger behavior. According to our analysis, we have realized that most existing trackbacks are blank-trackbacks that differ from the definition of weblog trackback. We will also discuss relationship between blog entries connected via trackback link.

1 Introduction

Recently, blogs have become widely used by general users as tools for putting out the information quickly and easily. According to a report[1] of Japanese Ministry of Internal Affairs and Communications (MIC) in May 2005, The cumulative number of bloggers (Internet users who maintain their blogs) in Japan is about 3.35 million (when considering bloggers who maintain two or more blogs, the net number of bloggers is about 1.65 million.) as of the end of March 2005. The MIC Study Group forecasts that by the end of March 2007, those numbers will increase to about 7.82 million and about 2.96 million, respectively. Thus, Web space becomes not only a place for getting information but also a place for

communication. It can be said that blogs have changed the way people use the Internet.

In Blogspace, general users can be not only contents consumers but also contents providers. We may say that blog contents are the mirror of public opinion. Consequently, it is reasonable to suppose that the importance of blog information is getting bigger.

In fact, some researchers do blog analysis such as blog community analysis and reputation analysis using blog data. It is known that the link structure among blog entries considerably influences the formation of blog communities on blogspace. Thus, it is very important to investigate hyperlinks and trackback links in order to understand characteristics of blog communities and blogger behavior. However, most researchers do not focus on trackback links, despite their importance in understanding the relations between blog entries. The reason why they do not focus on trackback is likely that it is not clear the significance and the meaning of trackback in Blogspace.

The "The Motive Internet Glossary [2]" says that

Trackback is a standard that can be used to automatically create a link between webpages (reciprocal link), usually between webpages on different websites.

Namely, A trackback is a function that can create a link from another blog page to user's own blog page independent of an intention of the another blog author. According to the above definition, trackbacks should exist with the opposite hyperlink in pairs. However, trackback links are automatically created by sending trackback ping, even if there are not the opposite hyperlinks. Actually, there exist such "blank-trackbacks" whose opposite hyperlinks are blank. Therefore, the purpose of this study is analyzing trackback usages to inspect weblog data quality for weblog mining and investigating its importance in understanding blogspace.

We first describe blogs and trackbacks, and related work. This is followed by a analyzing and considering of how trackbacks are used. We end with a summary and outline our plans for future work.

2 Blogs and Trackbacks

A blog entry, a primitive entity of blog content, typically has links to web pages or other blog entries, creating a conversational web through multiple blog sites.

Figure 1 shows a schematic of a typical blog site. A blog site is usually created by a single owner/blogger and consists of his or her blog entries, each of which usually has a permalink (URL: uniform resource locator) to enable direct access to the entry. Blog readers can discover bloggers' characteristics (e.g., their interests, role in the community, etc.) by browsing their past blog entries. If readers know the characteristics of a particular blog, they can expect similar characteristics to appear in future entries in that blog.

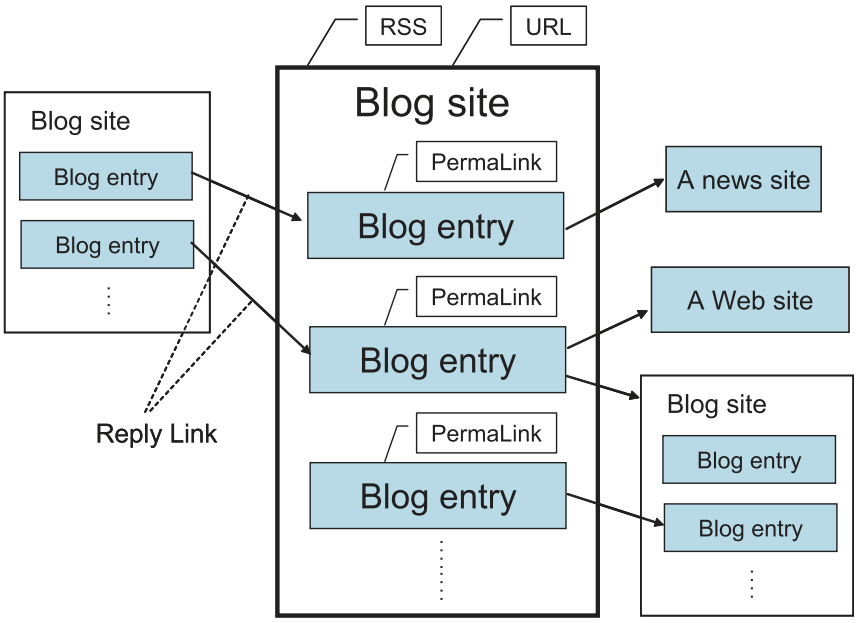

Fig. 1. Typical blog site

Figure 2 shows concept diagram of a typical trackback link.

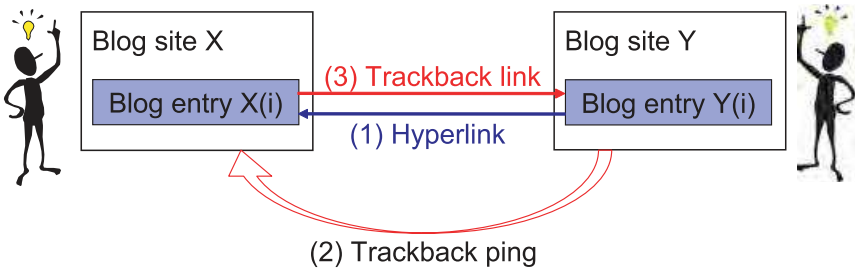

Fig. 2. Typical Trackback link with hyperlink

Originally, A trackback is a function that automatically create a link from another blog page to the user's own blog page when referring to another blog page. A typical procedure of creating a trackback is shown below:

(1) A user refers to a blog entry.

(2) The user send trackback ping to the referred blog entry.

(3) The referred blog system create a trackback link from the referred page to the referring page automatically.

Next, Figure 3 indicate concept diagram of a blank-trackback. As Figure 3 shows, trackback function automatically create a trackback link to when receiving trackback ping even if there is no the opposite hyperlink in pairs.

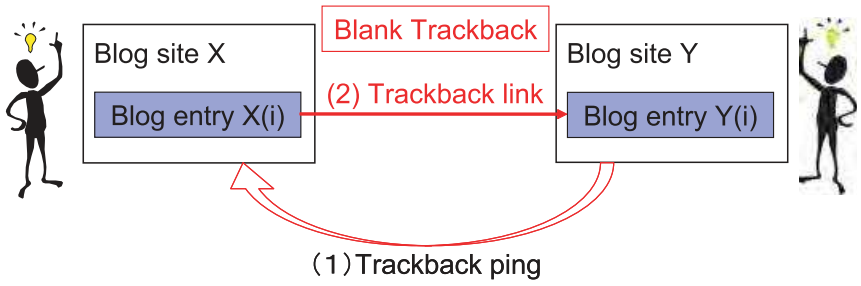

Fig. 3. blank-trackback

3 Related work

In related work on analyzing blogspace, Kumar et al. studied the burstiness of blogspace[3]. They examined 25,000 blog sites and 750,000 links to the sites. They focused on clusters of blogs connected via hyperlinks named blogspaces and investigated the extraction of blog communities and the evolution of the communities.

Gruhl et al. studied the diffusion of information through blogspace[4]. They examined 11,000 blog sites and 400,000 links in the sites, and tried to characterize macro topic-diffusion patterns in blogspaces and micro topic-diffusion patterns between blog entries. They also tried to model topic diffusion by means of criteria called Chatter and Spikes.

Adar et al. studied the implicit structure and dynamics of blogspace[5]. They also examined both the macro and micro behavior of blogspace. In particular, they focused on not only the explicit link structure but also the implicit routes of transmission for finding blogs that are sources of information.

Nakajima et al. studied how to discover important bloggers by analyzing blog threads[6]. They proposed a method of discovering bloggers who take important roles in conversations and characterized bloggers based on their roles in blog threads (a set of blog entries connected via usual hyperlinks). They considered that these bloggers are likely to be useful in identifying hot conversations.

However, their purposes were not to analyze trackback usages and to investigate its importance in understanding blogspace.

4 Analysis of Trackback Usages in Blogspace

4.1 Crawling through blog entries and extracting trackback links

The system crawls through RSS feeds registered on the RSS list and registers permalink of blog entries. Our RSS list have been created based on PING.BLOGGERS.JP[7] that open RSS feeds of JP domain to the public.

We need to extract the trackback links from html files of blog entries that have been crawled already. Therefore, we have to be able to recognize the scope

described the trackback data, based on an analysis of the HTML tag. However, each blog site server has its own tag structure so we need to set up rules for analyzing the tag structure of each blog site server that we want to analyze. By using the rules, we extract data of trackback links that are starting URL, destination URL and time stamp of the trackback link. Our target blog sites are limited to famous blog-hosting sites because naturally we are unable to set up rules for every blog site. We therefore set up rules for analyzing the tag structure of about 17 famous hosting sites of JP domain. We call them "blog sites for the analysis." We use 10,683,678 blog entries in "blog sites for the analysis" published from October 2005 to January 2006.

4.2 Relationship between entries connected via trackback links

Link-base relationship According to original definition of trackback, trackbacks should exist with the opposite hyperlink in pairs. We have investigated actual condition of link-base relationship between entries.

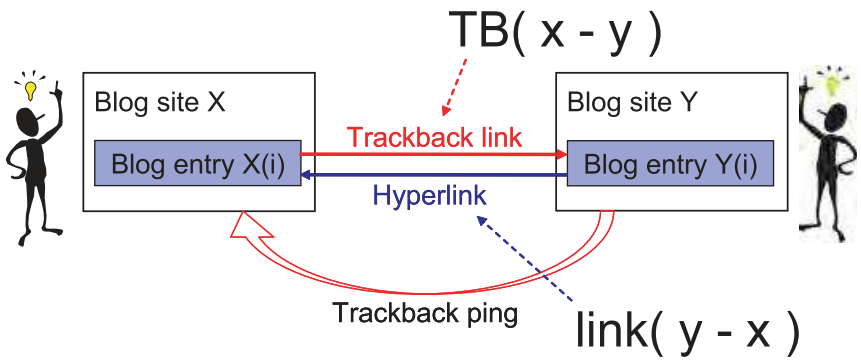

Fig. 4. Representation of a hyperlink and a trackback link

First, We examine the link-base relationship between entry(x) and entry(y) when existing trackback link (TB(x-y)) from entry(x) to entry(y). As indicated in Figure.4, TB(x-y) corresponds to a trackback link existing in entry(x), and it is created when receiving trackback ping from entry(y). link(y-x) corresponds to a hyperlink from entry(y) to entry(x).

Table. 1 shows link-base relationship of hyperlinks and trackback links. In this result, all patterns of existence of TB(y-x), link(x-y) and link(y-x) are investigated when a TB(x-y) exists. "link(x-y)O" denotes that link(x-y) exists. "link(x-y)X" denotes that link(x-y) does not exist.

In this case, blank-trackback means a situation that link(y-x) does not exist. It exists 99.08% of all patterns. In fact, almost all patterns in blog sites for the analysis are blank-trackback. Thus, the latest situations of trackback usages are different from the original definition of trackback.

Table 1. Link-base Relationship

	TB(y-x)O		TB(y-x)X	
	link(y-x) O	link(y-x) X	link(y-x) O	link(y-x) X
link(x-y)O	0.03%	0.29%	0.01%	0.76%
link(x-y)X	0.28%	11.28%	0.60%	86.75%

Moreover, another result that we focus on is the mutual blank-trackback relationship which is a situation that both TB(x-y) and TB (y-x) exist. This situation totally exists 11.88% of all patterns.

We may suppose that blank-trackbacks are kinds of spams, because a purpose of creating a blank-trackback may be to get more inlinks without making outlinks in order to get higher PageRank and more web users visiting own web page.

However, it is quite likely that bloggers give recognition each other in the mutual blank-trackback relationship. In this case, blank-trackback may not be spam.

Therefore, we are examining content-base relationships between entries that have blank-trackback relationship in the next chapter.

Content-base relationship In this section, we investigate content-base relationship between entries that have a blank-trackback link. The target data are 100 pairs of entries that have a blank-trackback. They are picked up at random from entries in the term from October 2005 to January 2006. The investigation is based on human judgment. The tester browse both contents of starting URL and destination URL of blank-trackback. According to trackback definition, a blogger who send a trackback ping should mentions in his/her blog entry about the content of blog entry that receive the trackback ping. Thus, the tester investigates which a blogger has mentioned about target blog entry of blank-trackback, or not.

Table.2 shows content-base relationship between entries that have blank-trackback. In all of the cases in Table.2, there exists TB(x-y).

Table 2. Content-base Relationship

	TB(y-x)O		TB(y-x)X	
	Mention(y-x) O	Mention(y-x) X	Mention(y-x) O	Mention(y-x) X
Related O	0%	7%	3%	57%
Related X	-	4%	-	29%

"Related O" or "Related X" denotes yes or no for existences of relation of the contents (Topics) between entries.

"Mention(y-x) O" or "Mention(y-x) X" denotes yes or no for existences of mentioning entry(x) in entry(y) that send trackback ping to entry(x).

As Table.2 indicates, there exist 33% of trackback links that point to unrelated entry. Actually, all of them are adult sites and spams indeed. Moreover, 4% in this 33% cases have mutual trackback relationship. In this case, all pairs of entries are mutual trackback-spams indeed.

We have recognized existences of trackback-spam. However, the percentage in number of blank-trackbacks is only 33%. Thus, we cannot say that blank-trackbacks are always spams.

In addition, there exists 57% of trackbacks which have no-mention but related to the opposite entry in pairs that is the target of trackback ping. Actually, they have the same topics. For example, review of movies and books, forecasting of horse racing, and so on. However, they do not mention contents of other entries at all. This kind of entries often has not only one way trackback but also mutual trackback. It seems that they form a "soft" blog community based on those trackback connection. Their connections are not strong and explicit.

4.3 Difference between usual hyperlinks and trackback links

Table. 3 shows a number of hyperlinks and a number of trackback links against total number of entries (10,683,678 entries) in "blog sites for the analysis." Now, we consider hyperlinks only appearing in the blog-description written by the blogger, except automatically-created links to access the past entries and commercial links, and of course except trackback links.

Table 3. Comparison between number of links and number of trackbacks

	Num. of Links	Num. of Links / Num. of Entries
Usual Hyperlink	4,613,581	0.432
Trackback link	1,696,177	0.159

As shown Table.3, an average number of hyperlinks per number of entries is 0.432 from October 2005 to January 2006, and an average number of trackbacks per number of entries is 0.159 in the same term. We can say that trackbacks become important to connect blog entries each other though the hyperlinks are still mainstream of connecting blog entries.

Figure.5 shows number of entries vs. the number of links contained in each entry. Figure.6 shows percentage of entries vs. the number of links contained in each entry.

As shown in Figure.5, most of entries containing hyperlinks (or trackback links) have less than 5-10 hyperlinks (or trackback links). The trends in the distribution between hyperlink and trackback link are similar though the number

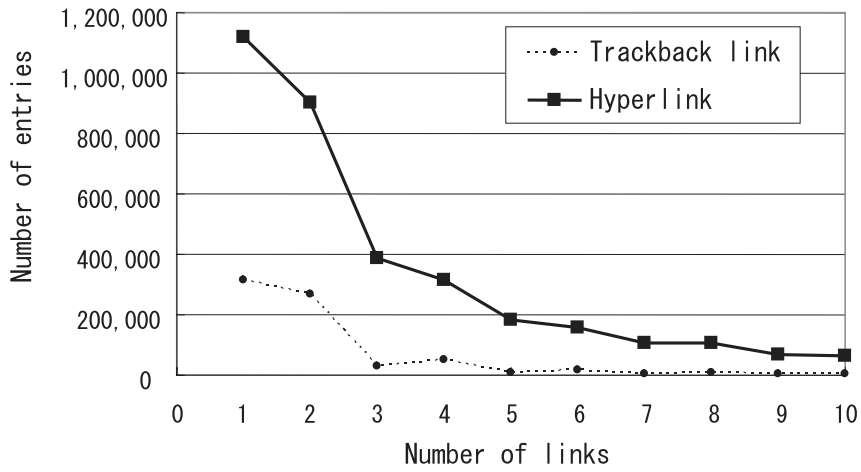

Fig. 5. Number of entries vs. the number of links contained in each entry

of hyperlinks is more than the number of trackback links. As shown in Figure.6, percentages of entries that have one or two trackback links are higher than the case of usual hyperlinks. It seems that connections via trackback links between blog entries are stronger than connections via usual hyperlinks because most of trackback links are often created by several particular bloggers.

Next, let's discuss difference between hyperlink-base blog communities and trackback-base blog communities. At beginning, we explain blog thread regarded as temporal blog community.

An example of a blog thread is shown in Fig.7. We define a blog thread as follows. A blog thread is composed of entries connected via links to a discussion among bloggers. Namely, a blog thread is a directed connected graph and is defined as follows.

$thread := (V, L)$
V is a set of blog entries.
$L \subseteq \{(e, e') | e \in V,\ e' \in V\}$
L corresponds to a set of links

Ideally, the entries in a blog thread should share common topics. The blog threads seem to be blog communities formed of blog entries that have similar topics via links.

Table.4 indicate numbers of hyperlink-base threads and trackback-base threads. In this investigation, we use blog entries published in October 2005 and November 2005 and their hyperlinks and trackback links. We can see the result in the cases that threads have more than 50 entries, more than 30 entries and more than 10 entries in Table.4.

As indicated in Table.4, the number of trackback-base threads is more than the number of hyperlink-base threads in all of three cases. As shown in Table.3, we may say that trackback links have about 3 times abilities to form blog threads

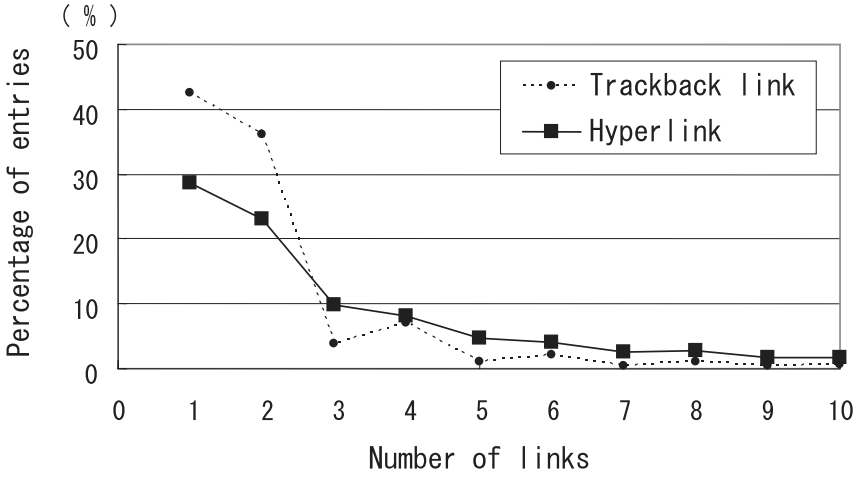

Fig. 6. Percentage of entries vs. the number of links contained in each entry

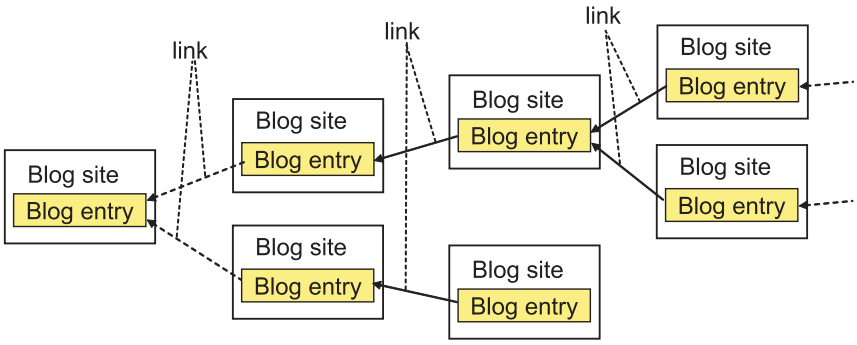

Fig. 7. Example of blog thread

as strong as hyperlinks because the number of trackback links is one third of the number of hyperlinks. Therefore, the trackback links are very important to investigate relationship between blog entries by analyzing blog communities.

4.4 Considerations

- **Differences between original definition and actual usage of trackback**
  As mentioned above, blank-trackback, whose opposite hyperlinks are blank, exists about 99% of all patterns. Moreover, blank-trackbacks are not always spams and they can form a "soft" blog community. It is different from original definition of trackback. It seems that we should re-define what trackback is, because such undefined usage becomes majority of trackback usages, at least for JP domain.
- **Existences of trackback spams**

Table 4. Comparison of number of threads between Hyperlinks and Trackbacks(Oct. 2005 and Nov. 2005)

	more than 50 entries	more than 30 entries	more than 10 entries
Num. of threads (via hyperlink)	51	145	1,032
Num. of threads (via trackback)	74	163	1,243

There exists 33% of trackback links that are spams indeed. It is quite likely that the trackback spams may become cause of great error when analyzing blog data. However, it is a problem of not only trackbacks but also Web itself. As Table.3 indicates, the trackback links are very important to form blog communities. Therefore, we had better consider trackbacks with developing spam filtering in order to analyze relationship between blog entries.

– **Blog communities based on trackbacks**
As mentioned above, trackback links have about 3 times abilities to form blog threads as strong as hyperlinks. The blog threads are regarded as temporal blog communities. Actually, characteristics of trackback-base communities are "soft" communities and a little different from characteristics of hyperlink-base communities. Accordingly, the trackback links are very important to discover and analyze blog communities.

5 Conclusions

In this study, we have analyzed trackback usages and investigating its importance in understanding blogspace, for understanding blogger behavior.

The results of this study can be summarized as follows:

– We have set up rules for analyzing the tag structure of famous hosting sites of JP domain and have analyzed trackback links
– We have investigated blank-trackback, whose opposite hyperlinks are blank, exists about 99% of all patterns.
– We have investigated that existences of "soft" blog communities based on blank-trackback connection.
– We have investigated difference between a number of hyperlinks and trackback links, and investigated ability to form blog threads. As a result, we have examined an importance of considering trackback when analyzing blog communities.

In addition, in future work we plan to investigate trackback data of blog entries in the other domain for understanding blogger behavior.

6 Acknowledgments

This research is partly supported by MEXT (Grant-in-Aid for Scientific Research on Priority Areas #19024058).

References

1. Analysis on Current Status of and Forecast on Blogs/SNSs, Press Release by Japanese Ministry of Internal Affairs and Communications on May 17th 2005C http://www.soumu.go.jp/joho_tsusin/eng/Releases/Telecommunications /news050517_2.html
2. The Motive Glossary - trackback http://www.motive.co.nz/glossary/trackback.php
3. R. Kumar, J. Novak, P. Raghavan, A. Tomkins: "On the Bursty Evolution of Blogspace", The Twelfth International World Wide Web Conference (2003). http://www2003.org/cdrom/papers/refereed/p477/p477-kumar/p477kumar.htm
4. D. Gruhl, R. Guha, D. Liben-Nowell, A. Tomkins: "Information Diffusion Through Blogspace", The Thirteenth International World Wide Web Conference (2004). http://www2004.org/proceedings/docs/1p491.pdf
5. E. Adar, L. Zhang: "Implicit Structure and Dynamics of Blogspace", WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004).
6. Shinsuke Nakajima, Junichi Tatemura, Yoshinori Hara, Katsumi Tanaka, and Shunsuke Uemura: "Identifying Agitators as Important Blogger based on Analyzing Blog Threads", Lecture Notes in Computer Science 3841, The Eighth Asia-Pacific Web Conference (APWeb2006), pp. 285-296, Springer-Verlag, (2006).
7. PING.BLOGGERS.JP, http://ping.bloggers.jp/