Discovering Important Bloggers based on Analyzing Blog Threads

Shinsuke Nakajima
National Institute of Information and Communications Technology
3-5 Hikari-dai Seika-cho
Soraku-gun Kyoto, Japan
snakajima@nict.go.jp

Junichi Tatemura
NEC Laboratories America,Inc
10080 North Wolfe Road,
Suite SW3-350, Cupertino, CA
95014, USA
tatemura@sv.nec-labs.com

Yoichiro Hino
Dept. of Social Informatics,
Kyoto University
Yoshida Honmachi, Sakyo-ku
Kyoto, Japan
hino@dl.kuis.kyoto-u.ac.jp

Yoshinori Hara
Internet Systems Research Laboratories, NEC Corporation
8916-47, Takayama-cho,
Ikoma, Nara, Japan
y-hara@da.jp.nec.com

Katsumi Tanaka
Dept. of Social Informatics,
Kyoto University
Yoshida Honmachi, Sakyo-ku
Kyoto, Japan
tanaka@dl.kuis.kyoto-u.ac.jp

ABSTRACT

A blog (weblog) lets people promptly publish content (such as comments) relating to other blogs through hyperlinks. This type of web content can be considered as a conversation rather than a collection of archived document. To capture 'hot' conversation topics from blogs and deliver them to users in a timely manner, we propose a method of discovering bloggers who take an important role in conversations. We characterize bloggers based on their roles in previous blog threads (a set of blog entries comprises a conversation),. We provide several definitions of bloggers' roles including (1) agitators who stimulate discussion, and (2) summarizers who provide summaries of the discussion. We consider that these bloggers are likely to be useful in identifying hot conversations. In this paper, we discuss models of blogs and blog thread data, and methods of extracting blog threads, discovering important bloggers, and acquiring important content from their entries.

Categories and Subject Descriptors

H.4.3 [**Information Systems**]: Information Systems ApplicationsCommunications Applications; H.2.8 [**Information Systems**]: Database ManagementDatabase Applications; J.4 [**Computer Applications**]: Social and behavioral sciences

General Terms

Human Factors

Keywords

blog thread, important blogger, agitator

WWW2005, May 10–14, 2005, Chiba, Japan.

1. INTRODUCTION

The broadband infrastructure and ubiquitous computing have created an environment in which people are continually online on the World Wide Web (WWW). Given this environment, people are increasingly publishing their reactions (e.g., comments and opinions) to current events (e.g., news). Users may state their opinion of a current news article, followed by other users who react to their opinions by stating a different opinion. In this sense, the web can be seen as a place for conversation rather than for archived documents. Triggered by an event, a hot conversation may quickly propagate from one site to another through the web.

A weblog, or blog for short, is a tool or web site that enables people to publish content promptly. The "Glossary of Internet Terms [1]" says that

"A blog is basically a journal that is available on the web. The activity of updating a blog is "blogging" and someone who keeps a blog is a "blogger." Blogs are typically updated daily using software that allows people with little or no technical background to update and maintain the blog. Postings on a blog are almost always arranged in chronological order with the most recent additions featured most prominently."

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.

Since conventional search engines treat the web as a snapshot of hyperlinked documents, they are not very effective for capturing conversational web content such as blogs. A new approach is required for timely delivery of hot conversations over multiple blogs on the web.

To capture potentially hot conversations on the web, we propose a method for discovering bloggers who take important roles in these conversations. This information on blog-

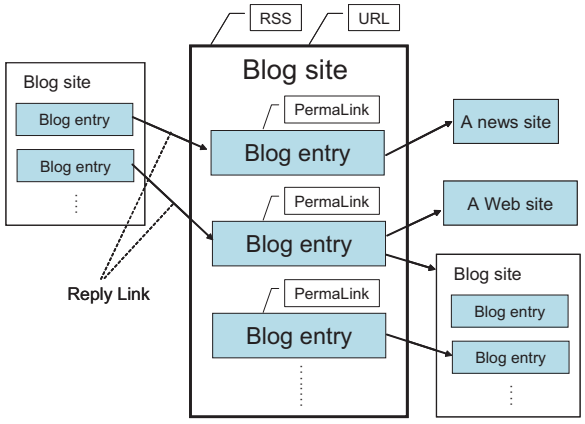

Figure 1: Typical blog site.

ger characteristics can then be used to acquire important hot conversations

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. Our goal is to develop a method of capturing hot conversations by automating readers' processes for characterizing and monitoring blogs.

In our method, an important blogger is defined on the basis of his or her role in a blog thread, i.e., a set of blog entries comprising a conversation on a specific topic. We think it is likely that bloggers take various roles in a thread, including acting as (1) an agitator who stimulates discussion, or (2) a summarizer who provides summaries of the discussion. We believe that these three types are important bloggers who are useful for identifying hot conversations.

We first describe related work and describe blogs and a model of blog thread data, and then the extraction of blog threads. This is followed by a discussion of how important bloggers are discovered, and examples of applications of our method. We end with a summary and outline our plans for future work.

2. RELATED WORK

In related work on analyzing blogspace, Kumar et al. studied the burstiness of blogspace[2]. 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[3]. 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[4]. 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.

However, their purpose was not to acquire important web content.

There are numerous reports of studies on topic detection and tracking[5]. In particular, there have been several studies on first story detection (FSD). The goal of FSD is to recognize when a new topic appears that has not been published previously. In this paper, we adopt a similar technique to discriminate agitators (aspect 3 (5.2)). Allan et al. studied first story detection[6]. According to this study, when a new story arrives, its feature set is compared to those of all past stories. If it is sufficiently different, the story is marked as a first story; otherwise, it is not.

Though these previous studies of FSD are relevant to us, we cannot adopt these methods directly since blog threads include relations between entries based on replylinks, which are different from a simple news stream.

3. BLOGS AND BLOG THREAD DATA MODEL

There are several definitions of blogs other than the one given in Section 1. Some are shown below.

- From "Web log:" A blog is basically a journal that is available on the web. The activity of updating a blog is "blogging" and someone who keeps a blog is a "blogger." [7]
- An easily updated personal website, generally updated daily and expressing opinions. [8]
- Web pages that are constantly updated with new commentary and links relating to a particular topic. Often very personal. [9]

That is, a blog is a website that anybody can easily update and use to express his/her own opinions in a public space. To put it another way, blogs are a storehouse of information that reflects public opinion. Although there is a lot of trivial information in blogspace, there is also a lot of important information.

Before defining our model of blog thread data , let us discuss these definitions of blog sites and blog entries. Examples are shown in Figures 2 and 3.

$$site = (siteURL, RSS, blogger^{+}, siteName, entry^{+})$$
$$entry = (permaLink, blogger, time, title^{?}, description, comment^{*})$$
$$comment = (blogger, permaLink, content, time)$$

A blog site has a site URL, RSS (really simple syndication), site name, and entries, and is managed by one or more bloggers. A blog entry has a permalink for access, a publication time, title, and entry description. A comment includes the content of the comment and the time when it was written. A blogger posts a blog to an entry identified by a permalink.

$$replyLink = (e_i, e_j), \quad (e_i \rightarrow e_j)$$
$$trackbackLink = (e_i, e_j), \quad (e_i \rightarrow e_j)$$
$$sourceLink = (e_i, w_i), \quad (e_i \rightarrow w_i)$$

where $e_i, e_j \in E$, E is a set of blog entries, and $w_i \in W$, W is a set of Web pages except blog entries.

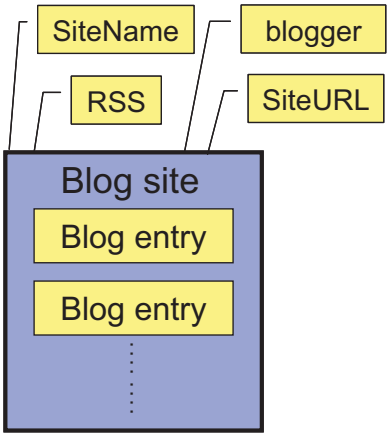

Figure 2: Example of blog site

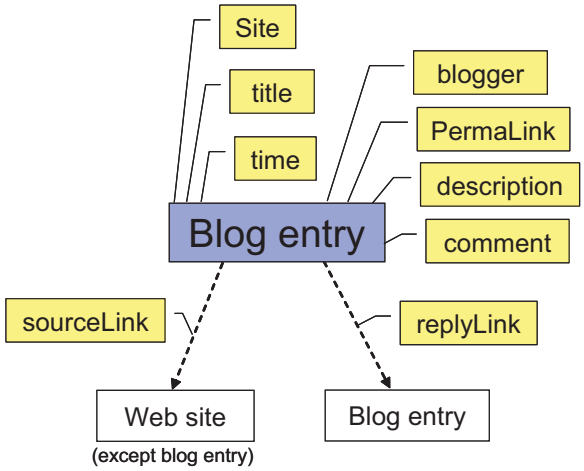

Figure 3: Example of blog entry

replyLinks and sourceLinks are hyperlinks described in the description of a blog entry to other blog entries or web pages. They do not include automatically added hyperlinks, such as links to previous entries, the next entry in the blog site, or other links unrelated to the content of the description of the blog entry. This is because we want to remove link noise to ensure that all blogspot.com pages point to blogger.com, etc. The method for removing link noise is described in Section 4.2. Here, a trackbackLink is a special case of a replyLink.

For $trackbackLink = (e_i, e_j)$, there is not only a replyLink of $(e_i \rightarrow e_j)$ but also a link of $(e_j \rightarrow e_i)$ to indicate the existence of a replyLink.

An example of a blog thread is shown in Fig.4. We define a blog thread as follows. A blog thread is composed of entries connected via replyLinks to a discussion among bloggers. There is one exception. As Fig. 4 indicates, sets of entries that are not connected to each other via replyLinks are regarded as being the same thread if they refer to the same website via a sourceLink. Comments attached a blog entry are not used in order to extract blog thread, because we want to identify important bloggers by analyzing blog thread so that it is not very important of comment author whose blog site could not be identified. Namely, a blog thread is a directed connected graph and is defined as follows.

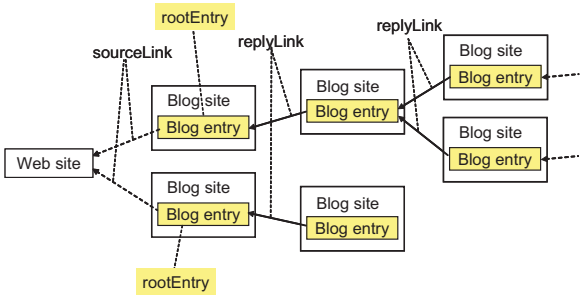

Figure 4: Example of blog thread

$thread := (V, L)$
$V = W \cup E, \quad L = L_s \cup L_r$
W is a set of websites. E is a set of blog entries.
$L_s \subseteq \{(e, e') | e \in E,\ e' \in W\}$
$L_r \subseteq \{(e, e') | e \in E,\ e' \in E\}$
L_s corresponds to a set of sourceLink.
L_r corresponds to a set of replyLink.

Ideally, the entries in a blog thread should share common topics. However, topics sometimes change at a particular point. In the future, we will separate a blog thread at that point even if both parts are connected via replyLinks.

4. EXTRACTION OF BLOG THREADS

This section shows how blog threads are extracted from real blog data.

4.1 Crawling through blog entries

First, our system crawls through blog entries to extract blog threads.

1. The system adds unregistered RSS feeds to the RSS list by crawling through opml files, "http://ping.bloggers.jp/opml.xml", "http://www.scripting.com/feeds/top100.opml" , and so on.

2. The system crawls through RSS feeds registered on the RSS list and registers the title, permalink, and list entry date as ungained entry data in the RSS.

3. Return to 1.

The RSS corresponds to the RDF Site Summary, which is actually an extension of RDF (resource description framework) language. RSS, which is an extensible metadata description and syndication format, is an XML application that conforms to the W3C's RDF specification. These days, most blog sites syndicate their content to subscribers by means of an RSS.

OPML (outline processor markup language) is an XML format for outlines. Originally developed as a native file format for an outliner application it has since been adopted for other uses, the most common being to exchange lists of RSS feeds between RSS aggregators.

Our system had registered about 400,000 RSS feeds (=blog sites) and over 10,000,000 entries as of January 15, 2005.

4.2 Extracting hyperlinks from descriptions of blog entries

We need to extract the hyperlinks from descriptions of blog entries to discover possible connections between the entry and other web pages (including blog entries). Therefore, we have to be able to recognize the scope of the description of an entry, 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. Our target blog sites are limited to famous blog-hosting sites and some famous bloggers' 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 25 famous hosting sites and some famous bloggers' sites. This enabled us to remove link noise from the replyLinks and sourceLinks.

The procedure for extracting hyperlinks from blog entries is given below.

1. The system crawls through the permalinks of entries that have not been crawled through in the entry list, and obtains the entries.
2. Descriptions of the entries are extracted from the HTML text by analyzing the tag structure.
3. Hyperlinks are extracted from the description and added to the list of links.

4.3 Extraction of blog threads

A blog thread is a set of entries connected to each other via replylinks and referring to a common web page via a sourcelink, as stated in section 3 (see Fig. 4).

The procedure for extracting blog threads is given below.

1. The system judges whether each hyperlink in the link list is a replyLink or sourcelink by checking whether the destination URL of the hyperlink appears in the entry list.
2. If it is a replyLink, the departure and destination URLs of the replyLink are checked to see whether or not they are registered in the existing thread data. If they are, they are added to the existing thread data. They become elements of a new thread if they do not.
3. If it is a sourceLink, the departure URL of the sourceLink is checked to see whether or not it is registered in the existing thread data, and the destination URL of the sourceLink is checked to see whether it is consistent with the Web page URL referred to by an entry registered in an existing thread. If there is an existing thread to be entered, it is added to the existing thread data. If not, it becomes an element of a new thread.
4. Return to 1.

The extracted thread data represents sets of entries, each with a date and link data. Consequently, the system can analyze the time-series data fir the entries in a thread (see Fig. 5) and the link structure of a blog thread (see Fig. 6).

In Fig. 6, each circle corresponds to a blog entry. each arrow corresponds to hyperlink between entries.

5. DISCOVERING IMPORTANT BLOGGERS

This section explains how to identify important bloggers.

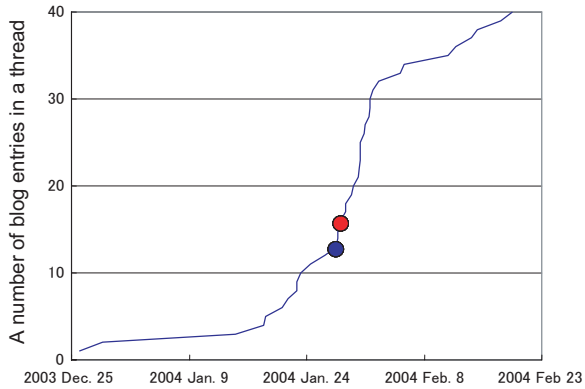

Figure 5: Example of time-series data of entries in a blog thread.

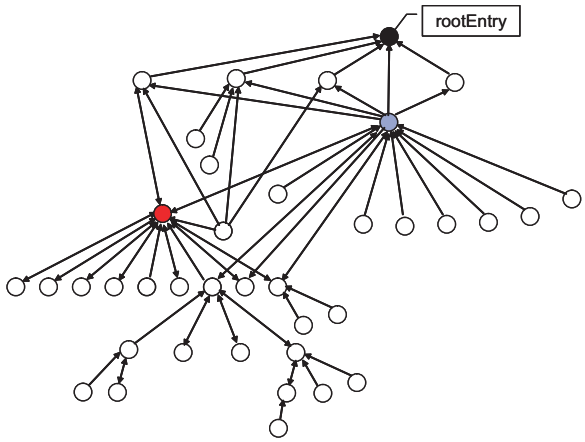

Figure 6: Example of link structure of a blog thread.

5.1 Important bloggers in blog threads

We define several potentially important bloggers in a conversation as follows:

1. **Agitator**

   An Agitator often stimulates the discussion in a blog thread so that it becomes more active. Thus, we may be able to predict whether a blog thread will grow by watching an Agitator's an entry. If the system statistically judges that threads often grow just after a particular blogger has published an entry, then that blogger is judged to be an Agitator.

2. **Summarizer**

   A Summarizer often publishes an entry that summarizes a hot blog thread by referring to many other entries. Thus, we may be able to obtain a summary of a blog thread that include hot topic by watching the entries of Summarizers. If the system statistically judges that a particular blogger publishes entries that often refer to other entries, then that blogger is judged to be a Summarizer.

In this paper, we focus on how to define and find an agitator and a summarizer, because we are interested in discovering important, hot blog conversations on a specific topic and

we believe that we can discover important blog conversation and its summary about a specific topic as blog threads by watching agitators and summarizers on specific topics.

5.2 Discriminants for Agitator

In this section, we introduce three aspects that characterize a blogger as an agitator. Given a blog thread, the following aspects discriminate an entry e_x from the other entries in a blog thread. We can then characterize the bloggers who publish such entries by aggregating the values from multiple blog threads.

- **Aspect 1: link-based discriminant**

  An entry by an agitator is characterized by the number of links to an entry from other entries. That is, an agitator is a blogger who is popular with other bloggers on a topic.

  e_x, an entry by an agitator is identified based on the following discriminant.

  $$(k_x) > \theta_1,$$

  where k_x is the number of entries in $thread_i$ that have a replylink to e_x.

- **Aspect 2: popularity-based discriminant**

  An entry by an agitator is characterized by a dramatic increase in the popularity of a thread shown by the number of entries published just after the agitators's entry.

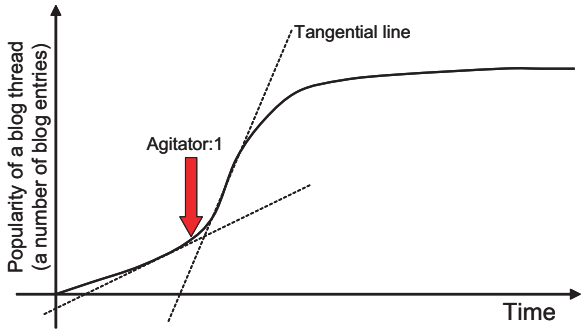

Figure 7: Feature of agitator in time-series data for popularity of blog thread.

  As shown in Fig. 7, the time-series data for the popularity of entries in a blog thread seem to reflect periods of stagnation and increased activity. Therefore, we consider that an agitator's entry appears between the end of a stagnant period and the beginning of a period of an increased activity.

  e_x, an entry by an agitator is identified using the following discriminant.

  $$(l_x/m_x) > \theta_2,$$

  where l_x is the number of entries in $thread_i$ that were published in the t (days) after e_x and m_x is the number of entries in $thread_i$ that were published in the t (days) before e_x.

  Therefore, (l_x/m_x) corresponds to an approximation of a second derivative value.

- **Aspect 3: topic-based discriminant**

  Entries by an agitator often have different characteristics in terms of content from entries in $thread_i$ that were published before e_x. In addition, they often have similar characteristics in terms of content to entries of $thread_i$ that were published after e_x. Namely, an agitator may have a big impact on a blog thread and may cause it to change.

  e_x, an entry by an agitator, is identified based on the following discriminant.

  $$\left\{ Similarity\left(\left(\frac{1}{n} \cdot \sum_{x+1}^{x+n} \vec{e}_i\right), \vec{e}_x \right) - Similarity\left(\vec{e}_x, \left(\frac{1}{n} \cdot \sum_{x-1}^{x-n} \vec{e}_i\right) \right) \right\} > \theta_3,$$

  where $\vec{e}_{x-n}$ is a feature vector of the n_{th} latest entry in $thread_i$ before e_x was published and
  $\vec{e}_{x+n}$ is a feature vector of the n_{th} earliest entry in $thread_i$ after e_x was published.

  In this paper, feature vectors are calculated by means of the TF(term frequency) values of the description of a blog entry. Similarity between feature vectors denotes text similarity calculated based on the cosine-correlation.

  Therefore, we believe that another discriminant for an agitator is represented by the point at which the topic of a blog thread changes.

We proposed three aspects for discriminating agitators. Though a single aspect is not enough to judge if an entry is by an agitator or not, using the three discriminants together should enable us to indentify agitators.

5.3 Discriminant for Summarizer

In this section, we introduce discriminant for Summarizer. Given a blog thread, the following discriminant identifies an entry e_x from the other entries in a blog thread. Consequently, we characterizes bloggers who published such entries by aggregating values from multiple blog threads.

- **link-based discriminant**

  An entry of summarizer is characterized by the number of links from an entry to other entries.

  e_x, an entry of summarizer, is identified based on the following discriminant.

  $$(p_x) > \theta_4,$$

  where p_x is the number of entries in $thread_i$ that have a replylink from e_x.

5.4 Examination of Discriminants for Discovering Agitators and Summarizers

We discuss the discriminants of aspects of agitator (5.2) and discriminant of summarizer (5.3) by means of examination of real blog data. Examples of real blog thread data are shown below. Fig. 8, 9 and 10 are data of thread (A). Fig. 11, 12 and 13 are data of thread (B).

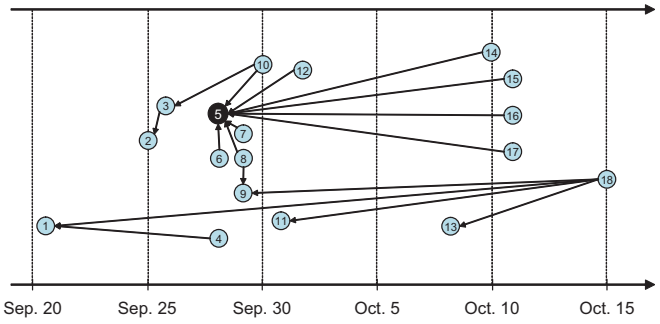

Figure 8: Example of link graph of a blog thread (A).

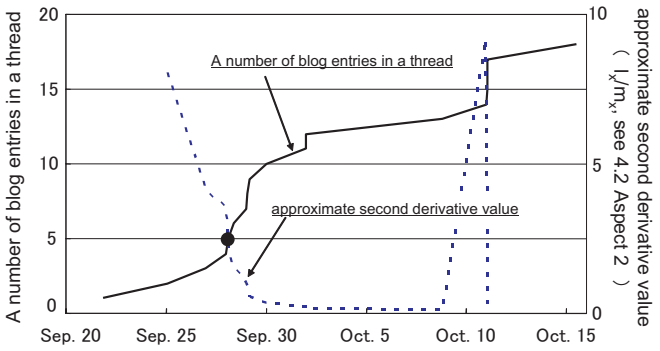

Figure 9: Time-series data of popularity and second derivative value of thread (A).

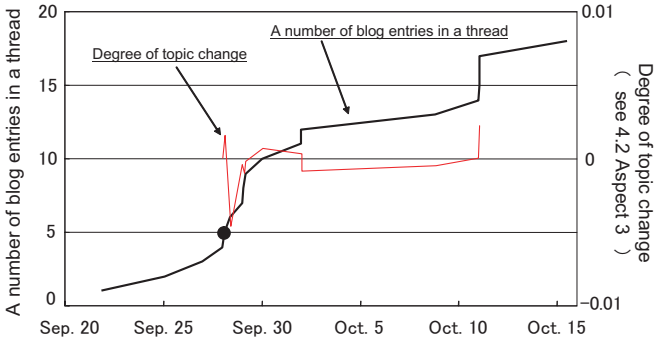

Figure 10: Time-series data of popularity and degree of topic-change of thread (A).

In Fig. 8 and Fig. 11, each circle corresponds to a blog entry. Each arrow corresponds to a replyLink between entries. The number in the circles denote the number from the oldest date entry to the newest date entry. The thread (A) has 18 entries, and the thread (B) has 34 entries. The

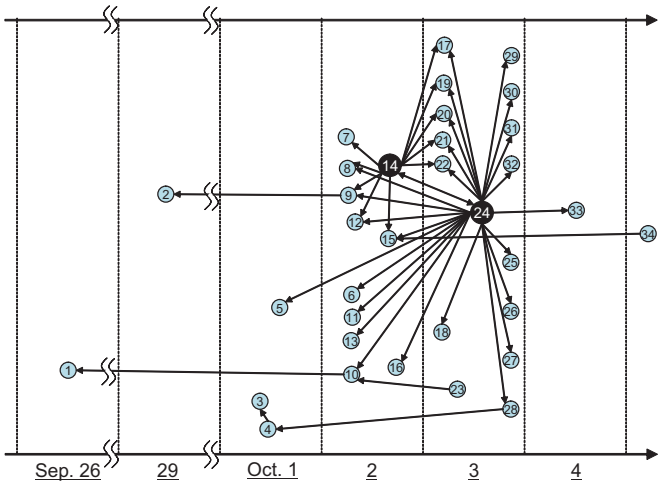

Figure 11: Example of link graph of the blog thread (B).

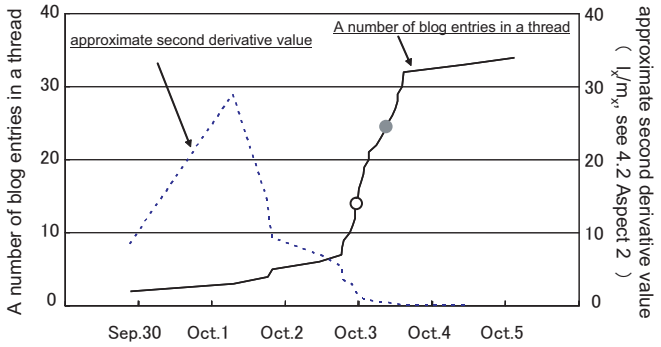

Figure 12: Time-series data of popularity and second derivative value of thread (B).

horizontal axis indicates publishing date of entries. Each date is in 2004. The vertical axis has no meaning.

There are some replyLinks from older entry to newer entry in Fig. 8 and Fig. 11. Generally, bloggers can edit their entries after publishing them. However, the date of entry are not changed. That is the reason why such replylinks exist. We allow such hyperlinks from newer entry to older entry as replylink.

As shown in Fig. 8, the entry of No. 5 of thread (A) seems something important in the thread, because it is cited from 9 other entries and entries in the thread increase just after published the entry of No. 5. Thus, the entry might be an entry of agitator based on aspect 1 of agitator. However, it is not always that many in-links mean good meanings. For example, some one may try to find fault with a blogger with referring via replyLink. Moreover, it is possible that an entry stimulate some one to write another entry nevertheless he/she does not refer to the entry. Thus, link analysis is not enough to judge agitator or not.

Next let us discuss discriminant of agitator and summarizer with time-series data of a blog thread shown in Fig. 9 The solid line in Fig. 9 corresponds to the time series data of popularity of the thread (A). The circle in black corre-

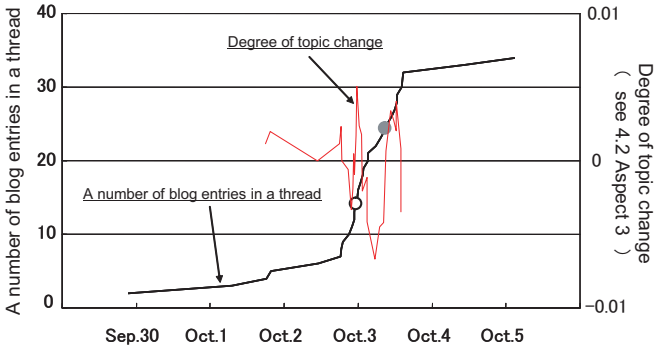

Figure 13: Time-series data of popularity and degree of topic-change of thread (B).

sponds to the entry of No. 5. The dashed line corresponds to (l_x/m_x), which is explained in Section 5.2. Namely, (l_x/m_x) corresponds to an approximation of a second derivative value of the time series data of popularity of blog entries.

As shown in Fig. 9, we can see that the values of (l_x/m_x) are high in the term just before the popularity is increasing drastically, at the point of No. 5 as well. Of course, there is no doubt about it, since (l_x/m_x) is a value of approximation of a second derivative of thread popularity. However, it is important to detect when the thread become hot by means of not human judgement but such objective value.

Next, we will discuss aspect 3 of agitator in the section 5.2. Fig. 10 shows the time-series data of popularity and degree of topic-change of the thread (A). The dashed line in Fig. 10 corresponds to the left part of the discriminant for aspect 3 of an agitator, as explained in Section 5.2. This corresponds to the degree of topic change. It is a fair possibility of changing topic of thread if the degree is high. According to Fig. 10, the degree of topic change become high at the No. 5 entry when just before the popularity is increasing drastically.

Hence, blogger of No. 5 entry of thread (A) seems to be candidate of agitator since the entry satisfies aspect 1, 2 and 3 of agitator. Acctually, the blogger is famous blogger about topic of IT trend whose blog site is

"http://blog.japan.cnet.com/umeda/".

In our system, we attempted to identify agitators by evaluating the three discriminants using real blog data. First, the left part of each discriminant is normalized the upper level of sum of these three are regarded as agitator.

$$S_A = \alpha \cdot Agi_1 + \beta \cdot Agi_2 + \gamma \cdot Agi_3$$
$$\alpha + \beta + \gamma = 100 \ (0 \leq S_A \leq 100)$$

S_A corresponds to the score for an agitator. Agi_1, Agi_2 and Agi_3 correspond to the average of the normalized left part of the discriminants for aspect 1, aspect 2, and aspect 3 of agitators. α, β and γ are weighting factors, and satisfy the following condition.

For another, let us discuss about thread (B).

As shown in Fig. 11, the entries of No. 14 and 24 of thread (B) also seem something important, because they are citing many other entries in the thread. Thus, they fairly satisfy discriminant of summarizer (5.3).

Next, let us examine the behavior of entries of No. 14 and 24, candidates of summarizer, in time-series data.

As same as Fig. 9, Fig. 12 shows the time-series data of popularity and second derivative value of thread (B). The solid line corresponds to time series data of popularity of the thread (B). The dashed line corresponds to an approximate second derivative value of time series data of popularity of blog entries. The circle in white corresponds to the entry of No. 14, the circle in gray corresponds to the entry of No. 24.

Unlike candidate of agitator, entries of No. 14 nad 24 have published not before the thread become active. Of course, it cannot be agitator if it have published before thread is hot, because great summarizer may summarize very hot part of a blog thread.

Further, Fig. 13 shows the time-series data of popularity and degree of the topic-change of the thread (B). The dashed line corresponds to a degree of topic change. From the figure, the topic of the thread has changed in term that thread has been more active. But it is hard to find another important relation between entries of summarizer candidates and the degree of the topic-change.

I found that these entries are described summary of a topic of horse racing predict by browsing these contents actually. Moreover, I believe that blogger of No.14 entry is a better summarizer than No.24 by browsing other entries in their blog sites.

Though we discuss about the examination of discovering candidates of agitator and summarizer, it is impossible to judge that a blogger is an important blogger or not, based on analyzing only one thread. Therefore, it is necessary of a statistical blog thread analysis in order to discover important bloggers, so that we should develop the system of discovering important blogger based on multiple blog threads analysis.

In addtion, we have to consider how to remove noises of blog data, since blog data has a lot of noises like miswritten html files, hyperlinks to non-existing url, advertisement links unrelated to the entry content, and so on.

6. EXAMPLE OF APPLICATION

This section shows a way of inserting important supplementary information with immediacy when browsing news content.

Though famous news sites and TV programs may deliver important information with immediacy, they cannot cover all the information relating to a news topic. We feel that it is important to provide a variety of supplementary information to audiences to avoid presenting them with biased opinions. For example, the information provided could include entries by important bloggers, part of a blog thread, related news referred to by an important blogger, and so on. Even when using a conventional technique such as a search engine, the system cannot pre-crawl supplementary information on a news item because the information may not have existed before the news event occurred. We therefore propose the following application shown in Fig. 14, which is the system that can insert important supplementary information with immediacy when a user browses news contents.

The operation of the application system is outlined below.

1. The system identifies important bloggers on particular

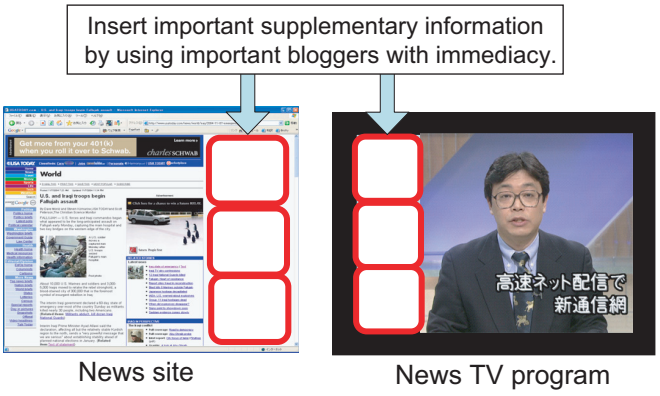

Figure 14: Example of applications inserting supplementary information with news contents.

topics.

2. When news contents is delivered, the system estimates the topic of the news content.
3. The system crawls the blog data from important bloggers related to the news content.
4. The crawled blog data are categorized using a clustering method.
5. The system provides blog data from important bloggers that differs to some degree from the news content, because data that is the same as news content is not useful as supplementary information.

In future work, we plan to work on identifying important bloggers for various topics, detecting topics in news content, crawling blog data from important bloggers, clustering blog data and presenting information that supplements news content.

7. CONCLUSIONS

we proposed a method for identifying important bloggers and acquiring important content from their blog entries. The results of this study can be summarized as follows:

1. A blog data model, blog site model, blog entry model, and blog thread model were defined.
2. We described a method of extracting blog threads, and extracted threads from real blog data (more than 10,000,000 entries) registered in our system.
3. We described a method for identifing agitators and summarizers as important bloggers by establishing discriminants for them.
4. We proposed a way of providing important supplementary information with immediacy when browsing news content.

In addition, in future work we plan to:

- Investigate a suitable method for detecting the topic of blog contents to identify the area of expertise of important bloggers.
- Develop the system of identifying important blogger based on multiple blog threads analysis.
- Design and develop a prototype system for providing important supplementary information with immediacy when browsing news content.

8. ACKNOWLEDGMENTS

This research is partly supported by the research for a Grant of Scientific Research (16016247 and 14208036) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and the Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of the Ministry of Education, Culture, Sports, Science and Technology, Japan).

9. REFERENCES

[1] Matisse's Glossary of Internet Terms http://www.matisse.net/files/glossary.html
[2] 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/p477-kumar.htm
[3] 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
[4] E. Adar, L. Zhang: "Implicit Structure and Dynamics of Blogspace", WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics (2004).
[5] J. Allan: "Topic Detection and Tracking", Kluwer Academic publishers (2002).
[6] J. Allan, V. Lavrenko, H. Jin: "First Story Detection In TDT Is Hard", In Ninth International Conference on Information Knowledge Management (CIKM'2000) (2000).
[7] MITSOL Help - Glossary http://www.mitsol.co.za/help_glossary.htm
[8] eGlossary http://www.internettime.com/itimegroup/eglossary.htm
[9] Glossary - Access eGovernment http://www.access-egov.info/glossary.cfm?xid=PA