

Analyzing Concerns of People from Weblog Articles

Tomohiro FUKUHARA¹, Toshihiro MURAYAMA¹ and Toyoaki NISHIDA²

¹Research Institute of Science and Technology for Society

{fukuhara, tmuraya}@ristex.jst.go.jp

²Department of Intelligence Science and Technology,

Graduate School of Informatics, Kyoto University

nishida@i.kyoto-u.ac.jp

Abstract

A system for analyzing concerns of people from Weblog articles is proposed. The system called Kanshin analyzes collective and personal concerns by collecting Weblog articles. The system collects RSS (RDF Site Summary) files of Japanese Weblog sites. The system provides keywords of the day and the month. Several patterns of collective and personal concerns are described.

1 Introduction

Understanding concerns of people is important for understanding a society. In the city of Pompeii which is one of cities of the Roman empire, various graffiti are found on walls¹. In the Roman era, there were various problems as well as today. From the graffiti, we can find personal and social concerns of people. Today, understanding concerns of people is important for solving social problems. There are many social problems in our society. For example, we have concerns over BSE (Bovine Spongiform Encephalopathy), SARS (Severe Acute Respiratory Syndrome), GMO (Genetically Modified Organism), and so on. For tackling with these problems, understanding concerns of people is important for finding key points of the problems to be solved.

We propose a system for understanding concerns of people from Weblog articles. Because Weblog has become one of important information channels to publish our thoughts and ideas on the Internet, we collect and analyze Weblog articles for finding concerns of people from collective and personal viewpoints. Users of this system can understand current concerns of people on his or her Web browser instantly.

This paper consists of following sections. In Section 2, we describe the aim of this research, and requirements for the prototype system. In Section 3, we describe an overview of the prototype system called *Kanshin*. In Section 4, we describe several patterns of social (collective) concerns found by the prototype system. In Section 5, we describe an approach to find *calm words* which have been topic-indicating words but mentioned rarely in recent articles, and some examples of calm words. In Section 6, we describe an approach and analysis results of personal concerns. In Section 7, we discuss differences between our system and other works. In Section 8, we summarize arguments of this paper, and describe the future work.

2 Analyzing concerns of people from Weblog articles

In this section, we describe (1) the aim of this research, and (2) requirements for the prototype system.

¹http://www.noctes-gallicanae.org/Pompeii/graffiti_1.htm (in French; accessed February 15, 2005)

2.1 The aim of this research

The aim of this research is to understand concerns of people from collective and personal viewpoints. This aim is related to the mission of our institute. The mission is to understand social problems, and propose solutions for them². According to this mission, we consider that gathering various thoughts and opinions of people on social problems is important for finding structure and key points of the problems.

So far, gathering thoughts and opinions massively was not easy. Although we often see TV and radio programs where audience send their thoughts and opinions to TV and radio stations via a telephone or e-mail or a facsimile, it is hard to keep continuing this kind of survey for a year. This kind of survey can be said *active survey* because one has to pay cost for gathering thoughts and opinions. A census can be seen as a kind of active survey.

In this paper, we focus on *passive survey* in which one observes streams of data, and analyze them. Today, Weblog (Blog), which is a writing style of Web articles sorted by reverse chronological order, becomes popular on the Internet. Many people called *Webloggers (bloggers)* publish articles on their Weblog sites day by day. A large number of articles can be collected. Furthermore, because most Weblog sites provide RSS (RDF Site Summary) file³ which is a small XML file (around 10k bytes) and containing an outline of a site, we can find updates of a site automatically by using a software called *RSS reader*. By collecting RSS files, we can find concerns of people on social problems.

2.2 Requirements

We consider that followings are needed for an analysis tool for understanding concerns of people, i.e., (1) collecting Weblog articles automatically, and (2) visualizing analysis results.

The first is to collect Weblog articles automatically. The system should collect and index articles so that users of this system can retrieve a number of articles effectively. Because a large number of articles are published day by day, indexing and retrieving articles effectively is important.

The second is to visualize analysis results. Facilitating users' understanding of analysis results is important. The system should visualize analysis results by converting text and numerical data into graphs so that users can understand the results intuitively.

3 Prototype system: Kanshin

In this section, we describe (1) the architecture, and (2) functions of the prototype system called *Kanshin*.

3.1 System architecture

The system consists of (1) a database, (2) a Web server, and (3) several Perl scripts for collecting and analyzing Weblog articles. Figure 1(a) shows an overview of the prototype system. The system collects RSS files from Japanese Weblog sites. An RSS file contains title, summary, date of publish, author, and category of an article. The system collects RSS files every 20 minutes. We started to collect RSS files since March 18, 2004. By now⁴, we collected 8,729,247 articles. About 20,000 articles are collected per day.

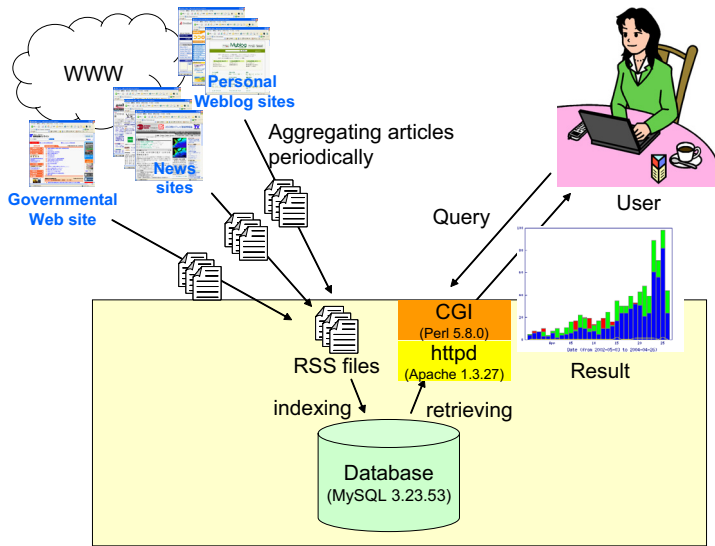
With respect to the internal architecture of the system, we use MySQL⁵ and several relational tables to manage words and articles. We use (1) *key_index* table which is an index table of keywords used for retrieving articles, (2) *term_kid* which is a table for listing articles containing a word whose id is *kid*, and (3) *rss_date* which is a table for storing articles for a date *date*. Definitions and samples of these tables are described in Appendix A.

²http://www.ristex.jp/english/mission/mission_e.html (accessed February 14, 2005)

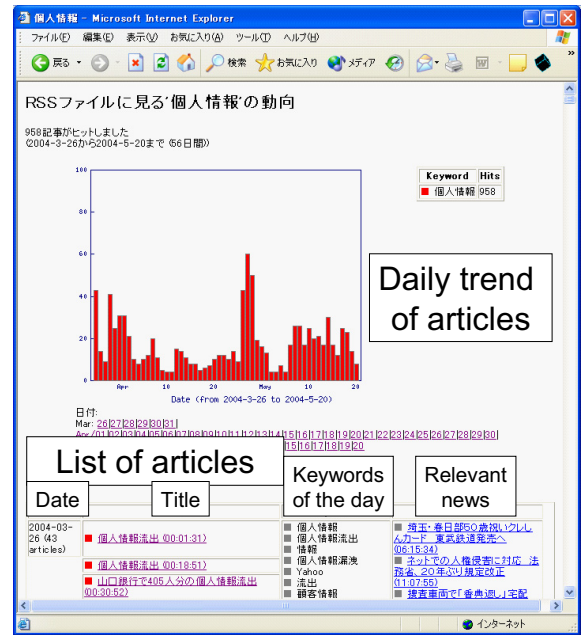
³<http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html> (accessed February 14, 2005)

⁴On February 14, 2004, 22:50:00 JST

⁵<http://www.mysql.org/> (accessed February 14, 2005)



(a) System architecture



(b) Screen image of the system

Figure 1. Overview of the prototype system.

RSS files are collected from (1) personal Weblog sites, (2) news sites, and (3) governmental Web sites. In case of the first and the second type of information, the system collects RSS files from Japanese Weblog ping servers such as Myblog Japan⁶ directly. In case of the third type of information, the system acquires Web pages from governmental Web sites, converts them into RSS files, and acquires information from those files.

Figure 1(b) shows a screen image of the system. The system accepts a query consisting of several keywords from users, and returns a graph which indicates daily trend of articles, and articles containing the keywords.

3.2 Functions

The system has following functions: (1) articles retrieval function, (2) finding relevant news articles function, and (3) finding daily and monthly topics function.

The first is article retrieval function. Users can retrieve Weblog articles by specifying a set of keywords connected by logical OR and AND operators. Figure 1(b) shows a screen image of search results. The figure contains a graph of daily trend of articles, and a list of articles. Users can find trend of the keywords.

Second, the system finds relevant news articles and displays them. News articles are retrieved by (1) keywords specified by the user, and (2) keywords extracted from articles.

Third, the system finds topic-indicating words called *daily topics* and *monthly topics* by calculating feature values of keywords. Table 1 shows an overview of the algorithm for finding monthly topics. In this algorithm, we count number of articles containing a word, and apply several thresholds. We use (30, 25, 0.6) for thresholds (τ_1, τ_2, τ_3).

Table 2 shows examples of monthly topics found by this algorithm. This table shows top three monthly topics for each month from April through November, 2004. '%' indicates the percentage of articles containing the word for each month. As

⁶<http://myblog.jp/> (in Japanese; accessed November 30, 2004)

Table 1. Algorithm for finding monthly topics.

1. Let \mathcal{M} be the set of months. If we want to know topics during Q months, $\mathcal{M} = \{m_1, m_2, \dots, m_Q\}$.
2. Let \mathcal{W} be the set of words appeared through Q months. If we find P words during Q months, $\mathcal{W} = \{w_1, w_2, \dots, w_P\}$.
3. For each $w_i (1 \leq i \leq P)$ in \mathcal{W} , repeat followings.
 - (a) For each $m_j (1 \leq j \leq Q)$ in \mathcal{M} , repeat followings.
 - i. Let a_{ij} be the number of articles containing w_i on m_j .
 - (b) Calculate $sum(a_i) = \sum_{j=1}^Q a_{ij}$, max value $max(a_i)$, and SD/average ratio $sd(a_i)/avg(a_i)$.
 - (c) Print w_i as a topic word of month m_j if $(sum(a_i) \geq \tau_1)$ and $(max(a_i) \geq \tau_2)$ and $(sd(a_i)/avg(a_i) \geq \tau_3)$.

Table 2. Monthly topics of 2004 (from April to November).

Month	Term (Japanese)	%	Month	Term (Japanese)	%
April	Iraq	1.6	August	Olympic games (Orinpikku)	2.4
	Cherry blossoms (Sakura)	1.5		Summer vacation (Natsuyasumi)	1.9
	Hostage (Hitojichi)	1.1		Player (Senshu)	1.8
	Release (Kaiho)	0.7		Competition (Taikai)	1.3
	Cherry blossom viewing (Hanami)	0.6		Athens	1.3
May	GW (GW)	1.5	September	Baseball team (Kyudan)	0.4
	Pension funds (Nenkin)	0.7		Entry (San-nyu)	0.3
	Golden Week	0.5		School term (Gakki)	0.2
	Unpaid (Mino)	0.4		Strike (Sutoraiki)	0.3
	Balley ball	0.4		Avoid (Kaihi)	0.2
June	The rainy season (Tsuyu)	0.9	October	Typhoon (Taifu)	4.0
	Kintetsu	0.4		Autumn (Aki)	2.0
	England	0.3		Earthquake (Jishin)	1.8
	Sulty (Mushi-atsui)	0.3		Cold (Samui)	1.7
	Portuguese	0.3		Niigata	1.2
July	Hot (Atsui)	4.7	November	Xmas (Kurisumasu)	1.0
	Summer (Natsu)	3.7		Next year (Rainen)	0.8
	Election (Senkyo)	0.8		Rakuten	0.6
	Publishing (Shuppan)	0.7		Autumn leaves (Koyo)	0.6
	The Star Festival (Tanabata)	0.6		Dragon quest (Dorakue)	0.4

shown in the table, monthly topics characterizing each month in 2004 are found⁷.

4 Patterns of social concerns

In this section, we describe several patterns of social concerns found by using the prototype system. We classified patterns into following categories: (1) *periodic pattern*, (2) *gradual increase pattern*, (3) *sensitive pattern*, (4) *trailing pattern*, and (5) *others*.

4.1 Periodic pattern

This pattern appears when there occur social events that are watched with keen interest periodically. The graph of periodic pattern shown in Figure 2(a) is a histogram of “Winter Sonata” which is the name of popular Korean TV drama broadcasted in Japan⁸. x axis indicates a date, and y axis indicates the number of articles. Because this drama attracted public attention

⁷Note that these words mainly characterize topics in Japan

⁸http://en.wikipedia.org/wiki/Winter_Sonata (accessed November 30, 2004)

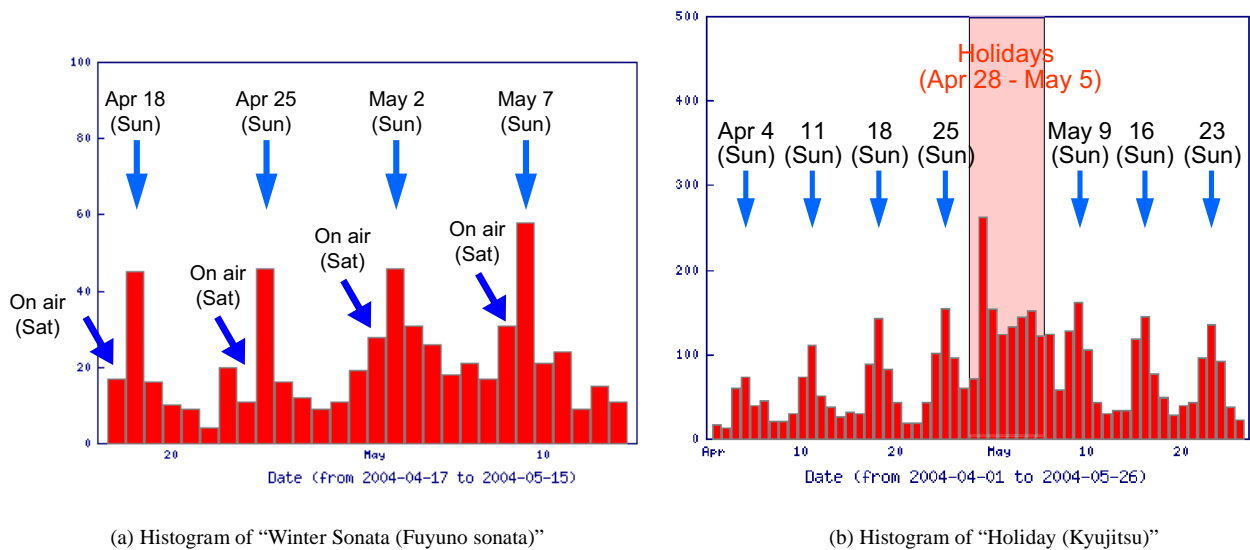


Figure 2. Examples of the periodic pattern.

in Japan, several periodic peaks appeared clearly in the graph. The drama was broadcasted on every late Saturday, so we can see periodic peaks on every Sunday. We can guess that bloggers write articles on this drama after they watched it from this graph. Other keywords of this pattern are “Holyday (kyujitsu)” (see Figure 2(b)), “Payday”, “the end of week (syumatsu)”, “the end of month (getsumatsu)”, “the end of year (nenmatsu)”, “Xmas”, “Summer vacation”, and so on.

4.2 Gradual increase pattern

In this pattern, a peak appears gradually. This pattern is also called *sleepier hit* in Glance et al. (2004). This pattern appears when people know a social event beforehand, and they have great interests on that event. Figure 3(a) shows a histogram of “GW (Golden Week)” which is the name of Japanese national holidays. x axis indicates a date, and y axis indicates the number of articles. As shown in the graph, people have strong interests in these holidays beforehand.

Other keywords of this pattern contain “Election”, “Typhoon”, “Summer”, and “Olympic games”⁹ (see Figure 3(b)).

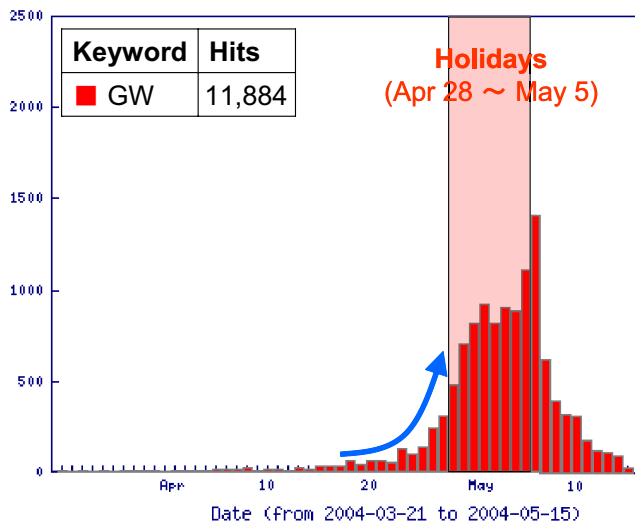
4.3 Sensitive pattern

This pattern appears when a serious matter, which has a heavy impact on the society, is broadcasted. This pattern is also called *Slashdot effect* (Adler (1999)). An example of the sensitive pattern is shown in the Figure 4. Figure 4(a) shows a histogram of “Winny” which is the name of a peer-to-peer (P2P) software¹⁰. x axis indicates a date, and y axis indicates number of articles. Articles on “Winny” were written intensively on May 10, 2004 when the programmer of this software was arrested.

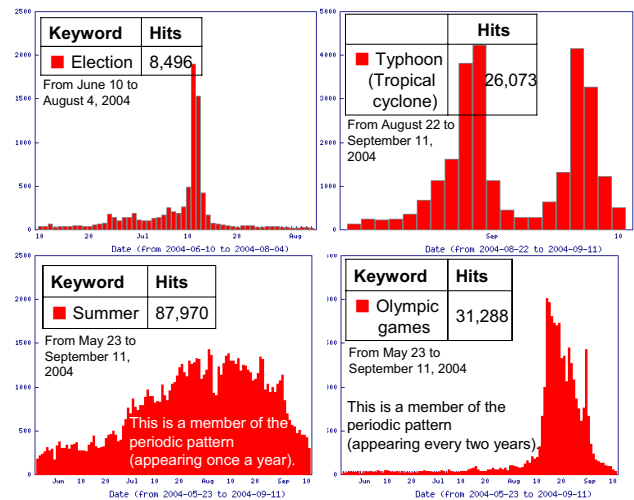
Another example of this pattern is “Earthquake” (see Figure 4(b)). The period of this figure begins from August 8 and ends with December 3, 2004. x axis indicates a date, and y axis indicates number of articles. During this period two big earthquakes hit Japan. On September 5, there was an earthquake off-shore southeast of the Kii peninsula (the Kii peninsula earthquake), and on October 23 there was an earthquake in the Niigata-Chuetsu region (the Niigata-Chuetsu earthquake).

⁹“Summer” and “Olympic games” are also members of periodic pattern, i.e., they appear periodically.

¹⁰<http://en.wikipedia.org/wiki/Winny> (accessed November 30, 2004)



(a) Histogram of "GW"



(b) Histograms of gradual increase patterns ("Election", "Typhoon", "Summer", and "Olympic games").

Figure 3. An example of the gradual increase pattern (Histogram of "GW").

The Number of articles exceeding 3,000 indicates the impact of these earthquakes because few words exceed one thousand except for general words such as "This" and "Yesterday" in this system.

Other keywords of this pattern contain words indicating an event or an accident that has a heavy impact on the society. "Nuclear power plant (Genpatsu)" showed this pattern when a heavy accident occurred in the Mihama nuclear power plant on August 9, 2004 (see Figure 6(b)).

4.4 Trailing pattern

In this pattern, concerns last after one or several issues occur. Figure 5(a) show an example of the trailing pattern. This figure is a histogram of "Iraq (Iraku in Katakana)". The period of this graph begins from April 1 to May 26, 2004 in which several Japanese people were kidnapped in Iraq. This accident was a very hot topic in Japan. TV and newspapers broadcasted this accident repeatedly. As well as mass media, various opinions on this issue appeared on Weblog sites. Figure 5(a) indicates that bloggers have great concerns on this accident. In the trailing part after an event or an accident happened, the focus on the event or the accident are changed. Figure 5(b) shows the change of focus on "Iraq". The period of the graph begins from March 21 to May 15, 2004. As shown in the graph, focus on Iraq was changed from "Hostage (Hitojichi)", "Release (Kaiho)", to "Abuse (Gyakutai)".

4.5 Others

This pattern is the rest case of the former patterns. Figure 6(a) shows a histogram of "Accident (Jiko)". Because so many accidents occur everyday, it is unable to distinguish specific concerns on an accident. This pattern appears when (1) a keyword is abstract such as "Accident", "Risk", and "Security", and (2) a keyword is not a hot topic at this time. Note that the latter type of keywords shows the sensitive and the gradual increase patterns when those keywords are paid attention by mass media or an influential Web site. For example, "Nuclear power plant (Genpatsu)" (see Figure 6(b)) was a member of this pattern until August 2004. However, it was changed to the sensitive pattern after a severe accident occurred in the Mihama nuclear power plant.

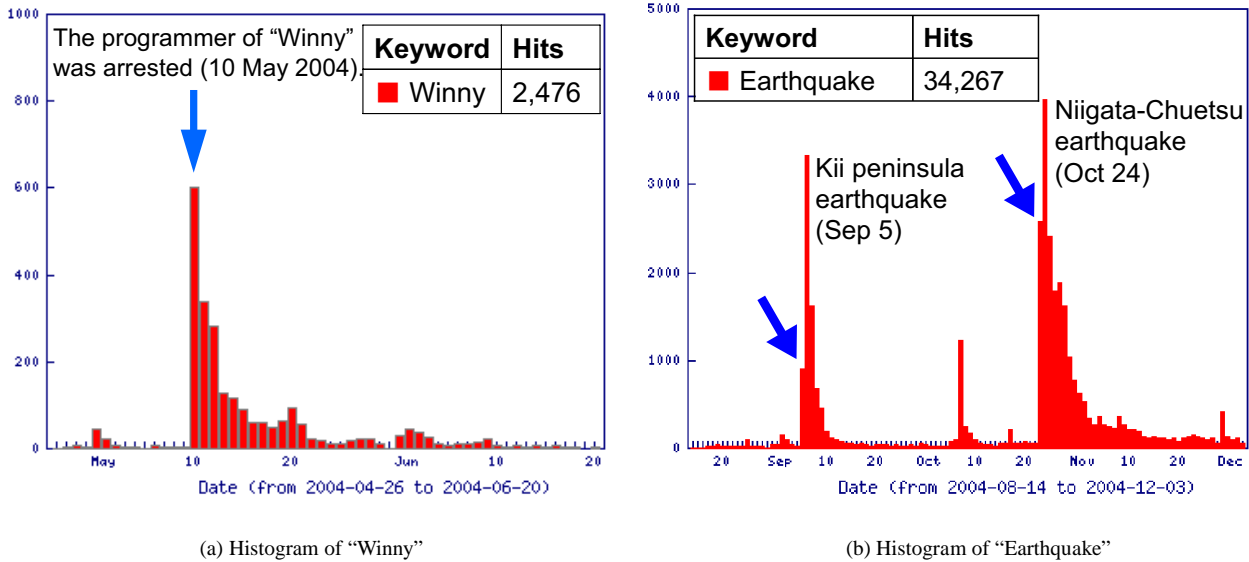


Figure 4. Examples of the sensitive pattern.

5 Calm words

In this section, we focus on *calm words* which have been hot topics once before but are not mentioned in recent Weblog articles. We describe (1) an approach to find calm words, and (2) the term for which bloggers remember topics.

5.1 Approach to find calm words

Calm words can be found by using (1) the timestamp of a word registered in the database within last N days (from $date1$ through $date2$), and (2) the average number m and standard deviation s of days N in which the word appears. For former criterion, we execute following SQL query on "key_index" table (see Appendix A.1).

```
SELECT * FROM key_index WHERE date >= 'date1' AND date <= 'date2';
```

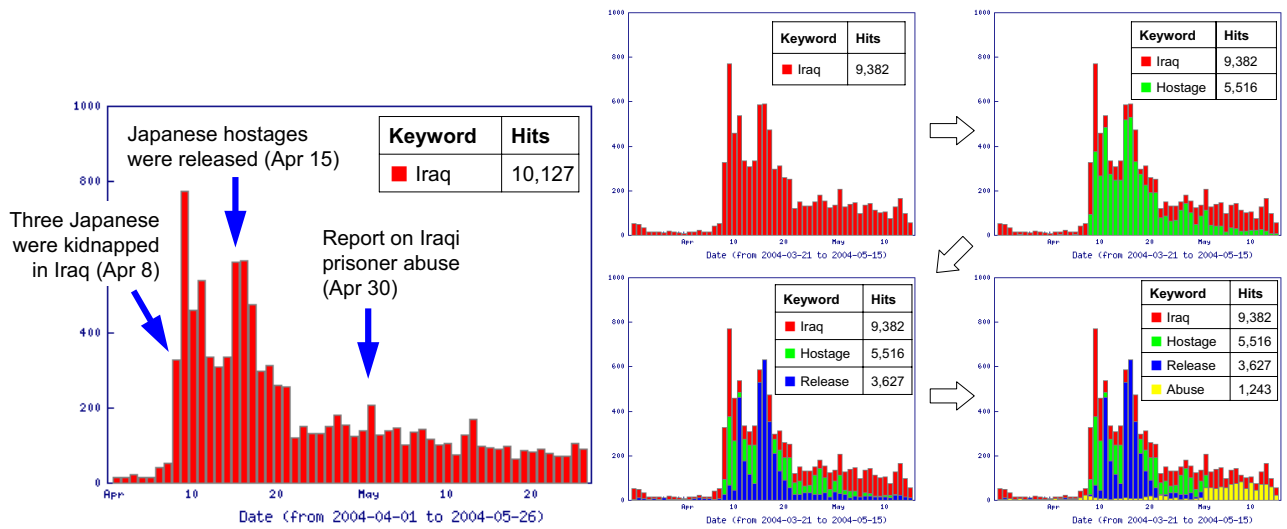
Then we calculate m and s for each word found by the above query. By applying $N = 30$, we found following words, i.e., "Häkkinen" ($m = 0.43$, $s = 1.05$) and "Berlusconi" ($m = 0.07$, $s = 0.25$) have not been mentioned since November 15, and "Armitage" ($m = 0.73$, $s = 2.22$) and "El Nino" ($m = 0.07$, $s = 0.25$) have not been mentioned since November 22¹¹.

Among calm words, we picked up some words indicating former topics. Table 3 shows words indicating former topics. These words were found by using the algorithm described in Table 1. In Table 3, we picked up words according to the following criteria.

1. A word that was former monthly topics is selected.
2. A word whose frequency is greater than 100 in "key_index" table (see Table 8) is selected.
3. A word that is a member of stop words such as "Blog" and "2004.11.08" is skipped.

We categorized words listed in the Table 3 into following categories.

¹¹Calm words are changed day by day because some words are updated and others are not in a day. These words are extracted on December 6, 2004, 12:00:00 JST.



(a) Histogram of "Iraq (Iraku in Katakana)"

(b) The change of focuses on "Iraq" from March 21 through May 15, 2004.

Figure 5. An example of the trailing pattern.

Seasonal topics

"April fool's day", "Cherry blossom (Yaezakura)", "Beginning of the rainy season (Tsuyuiiri)", "The end of the rainy season (Tsuyuake)", "Bonfire (Okuribi)", "Typhoon Tokage"

Summer related topics

"Burnt by the sun (Entenka)", "Suffer from the summer heat (Natsubate)", and "Lingering summer heat (Zansho)"

The name of holidays

"Golden week (Ogon shukan)", "G.W", "Day of the sea (Umi no hi)", "Respect for the aged day (Keiro no hi)", "Health sports day (Taiiku no hi)"

Sports and events related topics

"Silk Famous (Sirukufeimasu)", "U-23", "Radcliff", "Yokozawa", "Nakahata", "Bronze medal", "Rock odyssey", "Tenjin Matsuri Festival"

News topics

"The Republic of North Ossetia-Alania", "Marlon Brando", "Sasser", "Najaf"

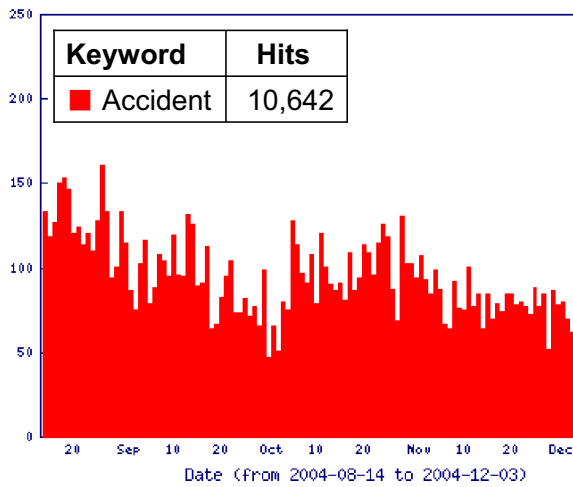
From this categorization, we found that bloggers rarely use words indicating former topics according to the number of days after an event or an accident happened. For example, "The Republic of North Ossetia-Alania" where the school siege was happened in September 2004 is rarely mentioned in Weblog articles (see Figure 7(a)). This might be affected by the collection of Weblog articles because we aimed to collect Japanese Weblog articles. If we collect articles written by people of the North Ossetia-Alania, the result might be different.

5.2 Remembrance of topics

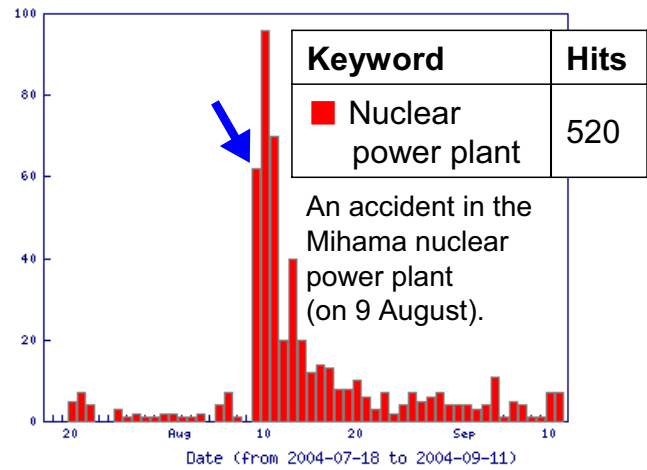
How long do bloggers remember former topics? As shown in Figure 7(a), few bloggers mention about former topics according to the number of days after an accident happened. For answering this question, we counted the number of words registered finally in the database.

Figure 7(b) shows a histogram of words registered finally in the database¹². We extracted words whose total frequency is

¹²This survey is conducted on December 7, 2004, 12:03:58 JST.



(a) Histogram of “Accident (Jiko)”



(b) Histogram of “Nuclear power plant (Genpatsu)”. This graph can be categorized into the sensitive pattern because the graph has a keen peak. Note that few people do not talk this topic except for several days when an accident happened.

Figure 6. Examples of the other pattern.

greater than 100. The period of the figure begins from November 17 and ends with December 7, 2004 (three weeks). Total number of words is 57, 161. Table 4 shows the number of words during this period.

As shown in Table 4, about 50.0% of words are updated by the day before (December 6). Then, percentage of words is decreasing according to the number of days from the final day. For example, 27.8% for December 5, 16.4% for December 4, 10.9% for December 3, and 7.8% for December 2 (see Table 4). This means that some words become obsolete day by day, and others are updated. From this result, we can guess that there are two types of words.

1. *Generic words* that are mentioned frequently such as “Today” and “They”.
2. *Topic indicating words* that are not mentioned usually but mentioned intensively in case of an event or an accident happened.

The former type of words keeps fresh because they are registered again in the database repeatedly. Meanwhile, the latter type of words is disappearing from bloggers’ memories according to the number of days after an accident happened. Note that seasonal words such as “Summer” become obsolete, but become active again in the next year.

6 Patterns of personal concerns

In this section, we describe preliminary analysis results of personal concerns. By collecting and analyzing Weblog articles of two persons, we found several common concerns between them.

6.1 Dataset

As a dataset, we collected articles written by Jun’ichiro Koizumi¹³ who is the prime minister of Japan, and Ryuichi Sakamoto¹⁴ who is a famous Japanese composer and musician. We collected 167 articles of Mr.Koizumi, and 51 articles of

¹³<http://www.kantei.go.jp/jp/m-magazine/backnumber/> (in Japanese; accessed 30 November 2004)

¹⁴<http://diary.nttdata.co.jp/> (in Japanese; accessed 30 November 2004)

Table 3. Example of calm words indicating former topics (extracted on 7 December 2004 21:34:00 JST)

Term (in Japanese)	Mean	SD	Date (Days)
Tenjin Matsuri Festival (Tenjin Matsuri)	0.03	0.18	8 Nov (-29)
Najaf	0.03	0.18	9 Nov (-28)
Silk Famous (Sirukufeimasu in Katakana)	0.17	0.58	10 Nov (-27)
Typhoon “Tokage” (Tokage)	0.07	0.18	11 Nov (-26)
The Republic of North Ossetia-Alania	0.03	0.18	12 Nov (-25)
Marlon Brando (Maron)	0.17	0.17	13 Nov (-24)
Bonfire (Okuribi)	0.10	0.30	14 Nov (-23)
U-23	0.13	0.43	18 Nov (-19)
Sasser (Sasser in Katakana)	0.10	0.40	19 Nov (-18)
G.W	0.03	0.18	20 Nov (-17)
Radcliff (Radokurifu in Katakana)	0.40	0.97	21 Nov (-16)
Yokozawa	0.10	0.30	22 Nov (-15)
April fool’s day	0.23	0.58	23 Nov (-14)
Nakahata	0.43	0.76	24 Nov (-13)
Rock odyssey	0.20	0.60	25 Nov (-12)
Cherry blossom (Yaezakura)	0.17	0.37	26 Nov (-11)
Golden week (Ogon Shukan)	0.07	0.24	27 Nov (-10)
Day of the sea (Umi no hi)	0.13	0.34	28 Nov (-9)
Beginning of the rainy season (Tsuyuiru)	0.17	0.37	29 Nov (-8)
Sasser (Sasser in English)	0.07	0.25	30 Nov (-7)
Suffer from the summer heat (Natsubate)	0.40	0.60	1 Dec (-6)
Burnt by the sun (Entenka)	0.60	0.92	2 Dec (-5)
Respect for the aged day (Keiro no hi)	0.27	0.51	3 Dec (-4)
Health sports day (Taiiku no hi)	0.37	0.66	4 Dec (-3)
Bronze medal	0.97	1.20	5 Dec (-2)
The end of the rainy season (Tsuyuuake)	0.23	0.56	6 Dec (-1)
Lingering summer heat (Zansho)	0.90	1.01	7 Dec (0)

Mr.Sakamoto. They updated Weblog articles once a week. Table 5 shows summary of the dataset. The periods of collections begins from May 29, 2001, and ends with November 25, 2004 for Mr.Koizumi. For Mr.Sakamoto, the period begins from August 31, 2003, and ends with September 26, 2004. Mean of words for Mr.Koizumi is 1081.9 and standard deviation (SD) is 1330.0. Mean of words for Mr.Sakamoto is 310.3, and SD is 241.7.

6.2 Approach to find personal concerns

We use *inverted document frequency* (*idf*) value of the vector space model ((Baeza-Yates, 1999, p.29)) for understanding personal concerns. The reason why we use *idf* value is that *idf* value of a word becomes smaller when many documents contain a specific word. We consider that words indicating concerns of a person are appeared in many articles, so we use *idf* value. For calculating idf_i value of an i -th word w_i , we use following formula.

$$idf_i = \log \frac{N}{n_i} \quad (1)$$

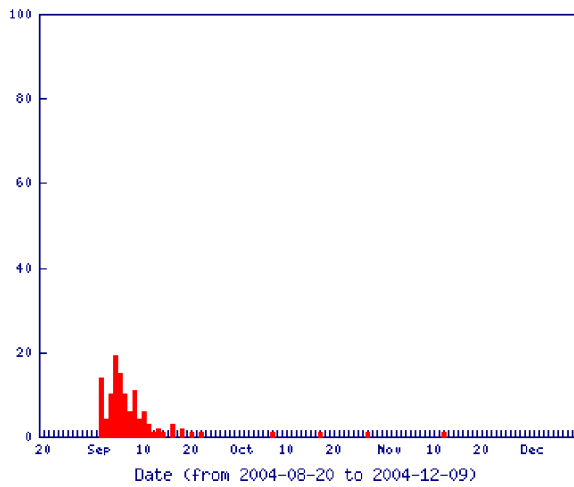
N is the total number of articles, and n_i is the number of articles in which word w_i appears in the formula (1).

6.3 Comparing concerns between persons

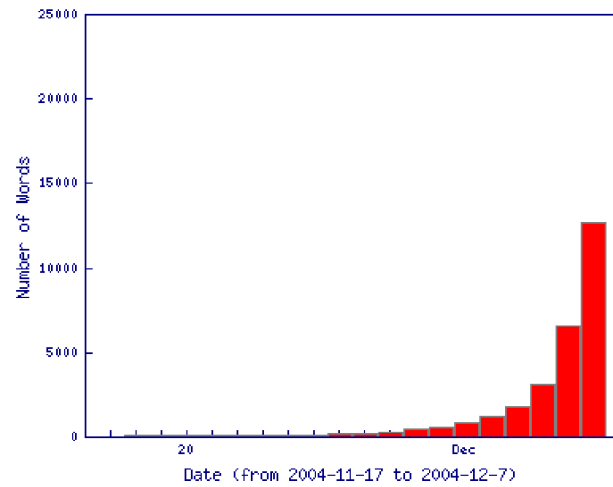
In this section, we first analyze concerns of Mr.Koizumi and Mr.Sakamoto, and compare their concerns.

Concerns of Mr.Koizumi

Table 6 shows top 20 words indicating concerns of Mr.Koizumi. This table is sorted by *idf* value. “Japan”, “Reform”, and “Issue” are extracted as words indicating concerns of Mr.Koizumi. Figure 8 shows a histogram for “Reform”. x axis shows



(a) Histogram of “The Republic of North Ossetia-Alania (Kita Ossethia Kyowakoku)” (on December 9, 2004, 22:30:00 JST).



(b) Histogram of words registered finally in the database (counted at December 7, 2004, 12:03:58 JST).

Figure 7. Behavior of calm words.

Table 4. Number of words finally registered in the database.

Date	Frequency	Amount	%	Date	Frequency	Amount	%
2004-11-17	25	25	0.04	2004-11-27	228	1090	1.91
2004-11-18	47	72	0.13	2004-11-28	296	1386	2.42
2004-11-19	48	120	0.21	2004-11-29	454	1840	3.22
2004-11-20	55	175	0.31	2004-11-30	604	2444	4.28
2004-11-21	67	242	0.42	2004-12-01	823	3267	5.72
2004-11-22	83	325	0.57	2004-12-02	1,194	4461	7.80
2004-11-23	99	424	0.74	2004-12-03	1,773	6234	10.91
2004-11-24	121	545	0.95	2004-12-04	3,143	9377	16.40
2004-11-25	133	678	1.19	2004-12-05	6,535	15,912	27.84
2004-11-26	184	862	1.51	2004-12-06	12,684	28,596	50.03
				2004-12-07	28,565	57,161	100.00

continue to the next table (↗)

the date, and y axis shows percentage of frequency of the word per number of words appeared in each article. “Reform” characterizes him very well because he always mentions about the reform of Japan.

Concerns of Mr.Sakamoto

Table 7 shows top 20 words indicating concerns of Mr.Sakamoto sorted by idf value. “Japan”, “Human”, and “Music” are extracted as words indicating his concerns. Figure 9 shows a histogram of “Music”. x axis shows the date, and y axis shows percentage of frequency of the word. Comparing with Mr.Koizumi (see Figure 8), he talks about “music” not so much, but he talks a lot when he talks.

Comparing concerns between Mr.Koizumi and Mr.Sakamoto

Comparing concerns of Mr.Koizumi and Mr.Sakamoto, “Japan”, “World”, and “Relation” are appeared commonly (see Table 6 and Table 7). “Government” can be seen as a common word although they use different notations, but their meanings

Table 5. Statistical data of articles.

Person	Articles	Days	Mean	SD
Mr.Koizumi	167	1,276	1,081.9	1,330.0
Mr.Sakamoto	51	392	310.3	241.7

Table 6. Words indicating concerns of Mr.Koizumi. Words with “*” are appeared in words of Mr.Sakamoto.

No.	Term (Japanese)	idf	Frequency
1	*Japan (Nihon)	0.22	545
2	Reform (Kaikaku)	0.46	529
3	Issue (Mondai)	0.65	291
4	Everyone (Minasan)	0.72	200
5	Economy (Keizai)	0.75	158
6	Nation (Kuni)	0.80	163
7	*World (Sekai)	0.80	145
8	Citizen (Kokumin)	0.87	142
9	Cooperation (Kyocho)	0.90	129
10	Everyone (Katagata)	0.91	117
11	Effort (Doryoku)	0.91	116
12	*Relation (Kankei)	0.96	134
13	Government (Seifu)	1.01	107
14	Society (Shakai)	1.06	119
15	Structure (Kozo)	1.13	82
16	International (Kokusai)	1.15	109
17	Heads of state (Shuno)	1.17	125
18	Safety (Anzen)	1.19	93
19	Diet (Kokkai)	1.19	87
20	Important (Juyo)	1.87	66

Table 7. Words indicating concerns of Mr.Sakamoto. Words with “*” are appeared in words of Mr.Koizumi.

No.	Term (Japanese)	idf	Frequency
1	*Japan (Nihon)	1.29	26
2	Human (Ningen)	1.53	13
3	Music (Ongaku)	1.73	24
4	Myself (Jibun)	1.85	13
5	Time (Jikan)	1.85	8
6	Children (Kodomo)	1.99	7
7	Sound (Oto)	2.14	15
8	America	2.14	11
9	Bush	2.14	9
10	Meaning (Imi)	2.14	9
11	*Relation (Kankei)	2.14	8
12	Today (Genzai)	2.14	7
13	*Government (Seiken)	2.14	6
14	*World (Sekai)	2.14	6
15	Words (Kotoba)	2.32	13
16	Cinema (Eiga)	2.32	12
17	Japanese (Nihongo)	2.32	9
18	Strange (Ijo)	2.32	6
19	Imagination (Sozo)	2.32	5
20	The Jomon (Jomon)	2.55	12

are similar (“Seifu” by Mr.Koizumi, and “Seiken” by Mr.Sakamoto). From this result, we can guess that they have common concerns on Japan government and the relationship between the world.

By following unique words appeared in Table 6 and Table 7, we found several characteristic words indicating the personality of a person. For example, “Music”, “Sound”, and “The Jomon” are characteristic words mentioned by Mr.Sakamoto. He also uses “America” and “Bush” because he talked about Iraq issues in his Weblog sites sometimes. “Strange (Ijo)” is also used in the context of global warming. From this result, Mr.Sakamoto has concerns over (1) the music, (2) international problems, and (3) global warming.

For Mr.Koizumi, “Reform”, “Economy”, “Society”, and “Diet” are characteristic words for him. “Heads of state (Shuno)” is also a characteristic word that indicates relationship between Japan and other countries. From this result, Mr.Koizumi has concerns over (1) the reform of Japanese social structures, (2) Japanese economic issues, and (3) relationship between countries.

7 Discussion

In this section, we discuss (1) the bias of concerns of bloggers, and (2) differences between related works.

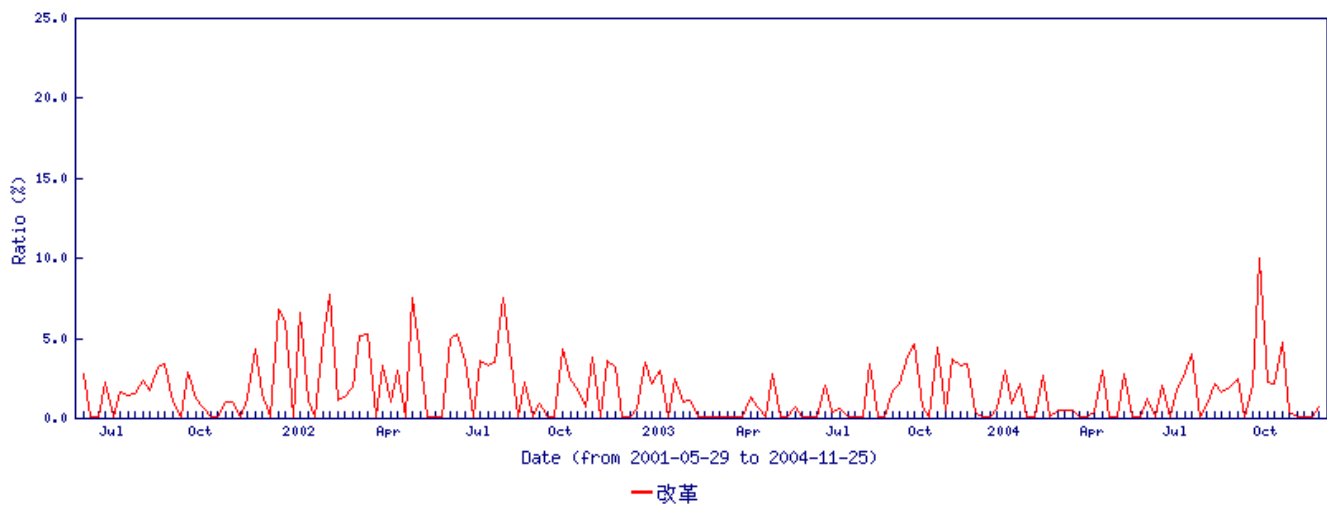


Figure 8. Personal concerns of Mr.Koizumi for a word “Reform”.

7.1 Bias of concerns of bloggers

Users of the Internet do not reflect the real world with respect to age, gender, occupations, and so on¹⁵. Furthermore, according to Miura and Yamashita (2004), 52.71% of bloggers are 20-30 years old, 29.68% are 30-40 years old, and 68.56% of bloggers are male¹⁶. Consequently, Weblog articles may not be adequate resources for conducting a rigid social survey.

Meanwhile, understanding social concerns from Weblog articles has another merit with respect to the speed. We aim to create a system with which a researcher who has concerns on social problems can find current hot topics instantly. By using the proposed system, researchers can find current topics instantly. This is an important feature of this system.

7.2 Related works

There are several related researches and tools such as blogPulse (Glance et al. (2004)) and blogWatcher (Nanno et al. (2004)). These services analyze Weblog articles, and provide histogram of words. One of differences is that our system analyzes concerns of people from collective and personal viewpoints. Understanding both of social and personal concerns is important for understanding social problems.

8 Conclusion and future work

In this paper, we described a system for understanding concerns of people by analyzing Weblog articles called Kanshin. By collecting Weblog articles, we can find concerns of people from collective and personal viewpoints. We described architecture of the prototype system, algorithms for finding monthly topics, and some examples patterns of social and personal concerns.

Our future work is to compare social concerns across countries by collecting Weblog articles described in various languages.

¹⁵http://www.cc.gatech.edu/gvu/user_surveys/ (accessed 30 November 2004)

¹⁶http://www.team1mile.com/asarin/research/04survey/frame_4.html (in Japanese; accessed December 10, 2004)

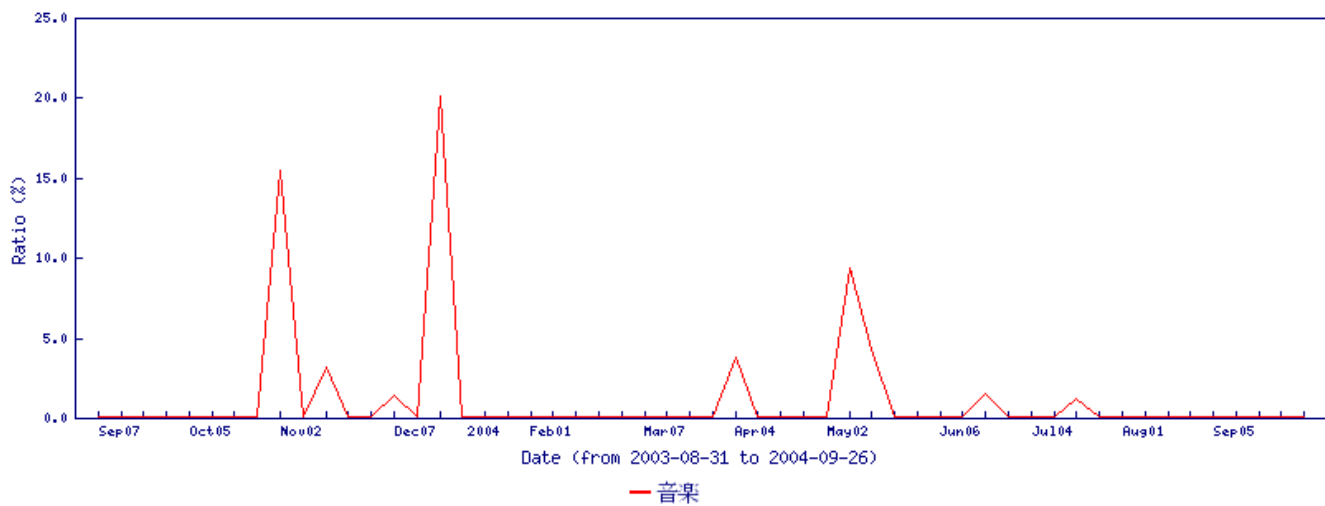


Figure 9. Personal concerns of Mr.Sakamoto for a word “Music”.

References

- Adler, S. (1999). The Slashdot effect. (online), available from (<http://ssadler.phy.bnl.gov/adler/SDE/SlashDotEffect.html>), (accessed 2004-12-10).
- Baeza-Yates, R. (1999). *Modern information retrieval*. Addison Wesley Longman Limited, U.K.
- Glance, N., Hurst, M., and Tomikiyo, T. (2004). BlogPulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. (available from <http://www.blogpulse.com/www2004-workshop.html>, accessed 2004-05-19).
- Miura, A. and Yamashita, K. (2004). Why do people publish weblogs?: An online survey of weblog authors in japan. In *Human Perspectives in the Ineternet Society: Culture, Psychology and Gender*, pages 43–50. WIT Press.
- Nanno, T., Suzuki, Y., Fujiki, T., and Okumura, M. (2004). Automatic collection and monitoring of japanese weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*. (available from <http://www.blogpulse.com/www2004-workshop.html>, accessed 2004-11-30).

Appendix A Definitions and samples of tables

We describe definitions and samples of tables used in the system. Tables listed in this section are dump lists of MySQL version 3.23.53-Max.

A.1 Table “key_index”

Definition and sample of “key_index” table are described in Table 8 and Table 9. Table “key_index” has following columns: (1) **kid**, (2) **term**, (3) **frq**, (4) **date**, and (5) **time**. **kid** indicates ID number of a word in the database. This column is associated with “term_kid” (see Table 10). **term** indicates label of the word. **frq** indicates the total number of a word. **date** and **time** indicate the date and time when this word is registered finally.

Table 8. Definition of “key_index”.

```

CREATE TABLE key_index (
  kid bigint(20) unsigned NOT NULL
    auto_increment,
  term varchar(80) NOT NULL,
  freq bigint(20) unsigned NOT NULL
    default '1',
  date date NOT NULL,
  time time NOT NULL,
  PRIMARY KEY (kid),
  KEY term_index (term),
  KEY date_index (date)
) TYPE=MyISAM;

```

Table 9. Example of “key_index”:
“key_index”.

kid	term	freq	date	time
1	Sunset	5984	2004-12-10	18:20:00
2	Hill	4790	2004-12-10	19:53:19
3	Night	787	2004-11-27	00:39:21
4	Year	8210	2004-12-10	20:38:58
5	Parents	1518	2004-12-10	19:46:11

A.2 Table “term_kid”

Table 10 and Table 11 show the definition and the sample of “term_kid”. Table “term_kid” has following columns: (1) **id**, (2) **num**, (3) **date**, and (4) **time**. **id** is the ID number of article which contains term *kid*. We can find this article by using “rss_date” table. **num** is the frequency of articles. **date** and **time** are date and time when this word is appeared.

Table 10. Table structure of “term_kid”.

```

CREATE TABLE term_kid (
  _id int(10) unsigned NOT NULL,
  num int(11) NOT NULL,
  date date NOT NULL,
  time time NOT NULL,
  KEY date_index (date),
  KEY rid_index (_id)
) TYPE=MyISAM;

```

Table 11. Example of “term_kid”: “term_3”.

date	_id	num	time
2004-06-29	16	1	14:06:49
2004-06-29	32	2	16:06:08
2004-06-29	68	1	23:24:03
2004-06-29	70	2	11:32:10
2004-07-01	312	1	18:56:37

A.3 Table “rss_date”

Definition and sample of table “rss_date” are described in Table 12 and Table 13. This table has following columns: (1) **id**, (2) **url**, (3) **title**, (4) **description**, and (5) **time**. **id** is local ID number of this article in this table. Articles are identified by **id** and **date** in our system. **url** is the URL of this article. **title** is the title of this article. **description** is the body of this article. **time** is the time of registration of this article.

Table 12. Table structure of “rss_date”.

```

CREATE TABLE rss_date (
  _id int(10) unsigned NOT NULL
    auto_increment,
  url varchar(255) NOT NULL,
  title text,
  description text,
  time time NOT NULL,
  PRIMARY KEY (_id),
  KEY title_index (title(80)),
  KEY url_index (url)
) TYPE=MyISAM;

```

Table 13. Example of “rss_date”:
“rss_20040629”.

_id	url	title	description	time
16	blog.l	Hakodate	Shinkansen Haya	14:06:49
32	www.do	Lunch 04031	After lunch, I	16:06:08
68	www.co	anne short	I was looking	23:24:03