Differences between Blogs and Web Diaries

Toshiaki FUJIKI
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
4259 Nagatsuta-cho Midori
Yokohama 226-8503 Japan
fujiki@lr.pi.titech.ac.jp

Tomoyuki NANNO
Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
4259 Nagatsuta-cho Midori
Yokohama 226-8503 Japan
nanno@lr.pi.titech.ac.jp

Manabu OKUMURA
Precision and Intelligence Laboratory
Tokyo Institute of Technology
4259 Nagatsuta-cho Midori
Yokohama 226-8503 Japan
oku@pi.titech.ac.jp

ABSTRACT

The number of weblogs (blogs) is rapidly increasing. However, there had been many 'web diaries' before the blogs arrived in Japan. 'Web diaries' are very similar to blogs, except that they are published without using blog publishing tools.

In this paper, we discuss whether the differences between blogs and 'web diaries' are significant for text mining or not. We conducted experiments to find the differences by comparing 'hot topic words' automatically extracted from blogs and 'web diaries'.

The results suggested that we could obtain better results by using blogs and 'web diaries' together rather than by only using the RSS data of blogs.

Categories and Subject Descriptors

H.5.4 [**Information Systems**]: Hypertext/Hypermedia; K.4.m [**Computers and Society**]: Miscellaneous

General Terms

Experimentation

Keywords

blog, 'web diary', content analysis

1. INTRODUCTION

Recently, the number of weblogs (blogs) has been increasing rapidly. Although the definition of blogs is not necessarily definite, blogs are generally understood to be personal web pages authored by a single individual and made up of a sequence of dated entries of the author's thoughts that are arranged chronologically. Blogs tend to be frequently updated and include links to other blogs. The content and purpose of blogs vary greatly, from links and commentaries about other web sites, to news about a companies/people, to diaries, photos, among others.

It is said that blogs date back to 1996, but they exploded in popularity during 1999 with the emergence of blogger (`http://www.blogger.com/`) and other easy-to-use publishing tools, such as 'Movable Type' (`http://www.movabletype.org/`). In 2002, a Newsweek article appeared estimating the number of weblogs to be half a million[4].

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

However, long before blog software became available in Japan, people wrote 'diaries' on the web (called 'web diaries'). These 'web diaries' were usually hand-edited or published with old-type CMS (Content Management System) software that did not support the ping server mechanism[8] and/or RSS distribution[1].

These 'web diaries' were quite similar to blogs in terms of their content, and people still write them without any blog software. There are numerous 'web diaries' throughout Japan, although most people now think of blogs as pages that are usually published with variants of blog publishing tools.

There has been some disagreement about whether the differences between blogs and 'web diaries' are meaningful or not. Although this is difficult to establish the differences, there have been some studies. Miura et al. analyzed weblog authors' thoughts[5][9]. They found that authors who thought of themselves writing a blog and those who thought of themselves writing a 'web diary' had different interests and this difference could be defined as the degree of blogging.

We are currently developing an automatic blog collecting and mining system called blogWatcher [6]. This system can recognize not only blogs but also 'web diaries' by analyzing the structure of web pages, and collecting these for mining. However, it has been controversial whether we should collect both blogs and 'web diaries', or we should collect only blogs, like other blog search engines, such as Technorati[1], Blogdex[2], and Daypop[3].[4]

In this paper, we define web pages published by blog publishing tools as blogs, and other web pages that look similar to blogs as 'web diaries'. We also discuss the differences between blogs and 'web diaries'. Although there may be many differences between blogs and 'web diaries', we focused on those that could influence the results of our text mining. We therefore automatically extracted words that described hot topics from blogs and 'web diaries', and analyzed them to find whether they were sufficiently similar to treat equally or not.

[1] `http://www.technorati.com/`
[2] `http://blogdex.media.mit.edu/`
[3] `http://www.daypop.com/`
[4] At `http://www.aripaparo.com/archive/000632.html`, the reader can find a useful list of blog search engines.

We will explain our method of collecting both blogs and 'web diaries' in the next section, and a method of extracting words representing hot topics in Section 3. In Section 4, we present the characteristics of data we used in the experiments. Section 5 describes our experiments and the results. Finally, we discuss the results and conclude the paper in Section 6.

2. COLLECTING BLOGS AND 'WEB DIARIES'

In this section, we will describe our method of collecting blogs and 'web diaries' together.

As previously mentioned, we defined 'web diaries' as web pages that looked similar to blogs, which were published without blog publishing tools. Therefore, to collect both blogs and 'web diaries', we could not rely on the metadata such as RSS and the ping mechanism, and we needed to identify blog/'web diary' pages from all the web pages by using the common format of the blog/'web diary' pages.

A page was judged to be a blog or a 'web diary' if and only if a sequence of entries that were articles for a day could be extracted from the page. The entries needed to satisfy the following constraints:

1. Entries needed to contain a date expression, which should have been at the top of the entry.
2. The date expressions for the sequence of entries needed to be consistently formatted[5], and arranged in ascending/descending order.
3. The tag sequence needed to be uniform for all the entries in the sequence.

The following three steps were needed to satisfy these constraints:

1. Extraction of date expressions,
2. Extraction of a sequence of dated entries,
3. Filtering non-blog pages.

For more details on the method, please refer to [7].

We use this method to develop the system called blogWatcher[6], which could collect blogs and 'web diaries'.

3. 'HOT TOPIC WORD' EXTRACTION

There have been few studies that have analyzed the content of blogs and/or 'web diaries'. One naive method to analyze the content of document sets is to measure the distance between the word vectors (i.e. Bag of Words representation) of the sets. However, we may not be able to obtain the effective results by applying this method to similar document sets.

Therefore, we automatically extracted words that represented the hot topics of document sets, and compared them. Differences in the 'hot topic word' of each document set mean the interests of writers/readers were different in each document set.

We used the burst identification algorithm we previously proposed[2] to extract 'hot topic words' automatically. Some

[5]'2003/1/2' and '2-Jan-2003' were considered to be inconsistent.

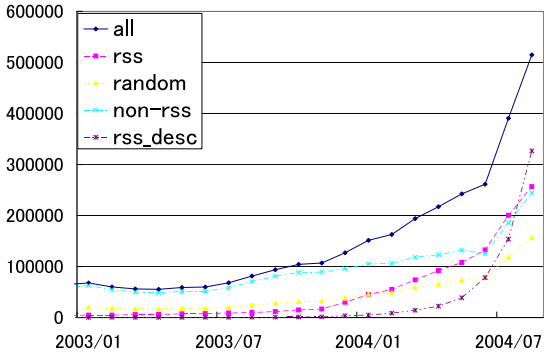

Figure 1: Number of entries in each dataset

burst identification algorithm indicated sharp rises in frequency as the topic emerged. In Kleinberg's original work[3], which targeted E-mail and research papers, his text mining algorithm tried to identify bursts in document streams in which messages arrived in temporal order. Simply put, a burst can be found by searching periods when the word tends to appear at shorter than usual intervals.

Unfortunately, as Kleinberg's original algorithm could not be applied to blogs, we extended the algorithm so that it could discover dense periods of 'burstiness' in the blogs[2]. The reason why the original algorithm could not be applied to blogs is that since the distribution of the blog entries was not uniform, the interval during which entries arrived tended to be shorter when more blog entries arrive, and consequently, more bursts tended to be erroneously identified. The experiments where we applyed our method to blogs revealed that it could extract 'hot topic words' accurately.

Table 1 lists the 'hot topic words' for Apr. 2004, automatically extracted with our method. These 'hot topic words' were extracted from the dataset 'all' (to be described later),

Table 1: Example 'hot topic words' (dataset all: Apr. 2004)

1	Hostage (in Iraq)
2	April Fool's Day
3	Japanese (rel. Iraq hostage)
4	Mitsuteru (famous comic artist)
5	Golden Week (consecutive holidays)
6	Greenery Day (national holiday)
7	Aljazeera (rel. Iraq)
8	Takato (rel. Iraq hostage)
9	seeing cherry blossoms
10	Uekusa (famous economist)

4. DATA SET

It is difficult to discriminate between blogs and 'web diaries' apparently. Therefore, we defined web pages published with blog publishing tools, which delivered RSS, as blogs, and web pages that looked similar to the former, published without blog publishing tools, as 'web diaries'. Since we cannot know whether there is a difference in content, this definition was based on functional differences.

We used the data collected with our blog collecting system (blogWatcher). As previously described, the system collected both blogs and 'web diaries'. Therefore, we had

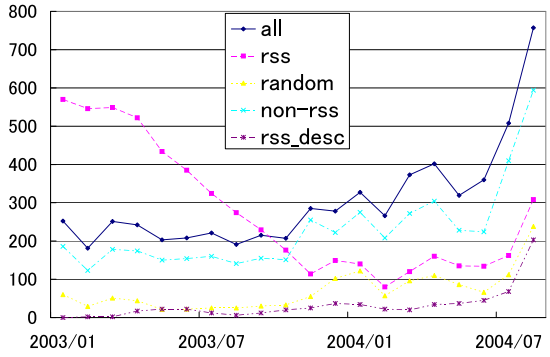

Figure 2: Number of 'hot topic words' extracted from each dataset

to distinguish blogs from 'web diaries'. We used the RSS Auto Discovery technique[6]. We could establish whether web pages corresponded to RSS or not with this technique, and distinguish blogs from 'web diaries'.

We collected blogs and 'web diaries' that had been published from Jan. 2003 to Sep. 2004, and prepared five datasets for the experiments, which we will describe in the next section.

1. all
   Includes blogs and 'web diaries' without discrimination. Corresponds to the integration of dataset 'rss' and dataset 'non-rss'.

2. non-rss
   Only includes 'web diaries' defined above.

3. rss
   Only includes blogs defined above. This dataset contains the full text of entries in contrast to 'rss-desc'.

4. random
   Includes blogs and 'web diaries', the same as 'all'. However, the number of entries included has been adjusted to be approximately the same as 'rss' by random sampling.

5. rss-desc
   Different to the above four datasets, we collected the RSS for this dataset independently. This means that the pages included in this dataset cannot always be included in the 'rss' dataset.

   This dataset uses description elements of an RSS as entry texts. A description element is a kind of summary of an entry and its value is derived from the lead sentences of the entry in many cases. Almost all of search engines for blogs treat these elements as the body text for entries, and index these elements for text searching.

Figure 1 plots the number of entries which each dataset contains. As you can see, the number of entries for 'rss' and 'rss-desc' is increasing rapidly, while those for 'non-rss' are stable. We can see from the figure that blog publishing tools are commonly used nowadays. The figure also shows that there were few entries for 'rss-desc' in the past. The reason for this is that RSS only contains the entries that have recently been updated. This means that we cannot obtain the old entries using RSS or the ping server mechanism.

Figure 2 plots the number of 'hot topic words' for each dataset. There are many 'hot topic words' for 'rss' for early 2003. This is caused by the duplicate entry problem (collecting the same entry from multiple URLs). This means that the period with few entries is influenced too much by the duplicate entries.

We used 'hot topic words' for each dataset extracted from May 2004 to Aug. 2004 for the experiments we describe in the next section.

[6]`http://diveintomark.org/archives/2002/05/30/rss_autodiscovery`

5. EXPERIMENTS

In this section, we will describe five experiments. One of the authors of this paper made subjective decisions for each evaluation.

5.1 Accuracy of 'Hot topic words' (Loose Standard of Assessment)

We first conducted an experiment to find the precision of 'hot topic words' extracted from each dataset. We determined whether the top 30 words represented the topic of the month or not. When the result obtained from a dataset is accurate enough, this means that the dataset contains many true topic words, and reflects the state of the real world. This also means that the dataset has no serious deviations.

However, it is difficult to determine whether the 'hot topic words' are true or not. Therefore, in this first analysis, we only assessed words as inappropriate if they were related to the events of other months or they had been extracted due to dataset deviations.

The results of this experiment are listed in Table 2.

Table 2: Accuracy of 'hot topic words' (Loose Standard of Assessment)

	all	non-rss	rss	random	rss-desc
May	100%	97%	97%	100%	93%
Jun.	97%	93%	90%	90%	87%
Jul.	100%	97%	93%	100%	97%
Aug.	100%	97%	100%	100%	87%
Avr.	99%	96%	95%	98%	91%

As you can see from the table, the results from four datasets were very accurate with the loose standard of assessment we used in this experiment. This means none of the datasets had any serious deviations, and could be used to extract appropriate 'hot topic words'.

However, the result obtained from 'rss-desc' was significantly less accurate than the other datasets (using chi-square test at 5% level).

5.2 Precision of 'Hot topic words' (Strict Standard of Assessment)

The results of the previous experiments showed that we could obtain accurate results from most of the datasets. However, many words that the evaluator did not know were judged to be correct with this standard. For example, a sports player who the evaluator felt was not that famous was judged appropriate for the 'hot topic words'. Although

we did not want to judge these words as being true, the evaluator did not know whether these words were really well known or not.

Consequently, we conducted a second experiment with a stricter judging standard. With this standard, we judged as appropriate words which were parts of phrases that represent true hot topics, and judged as inappropriate words which the evaluator did not hear about on newsmedia, such as names of non-popular sports players or racehorses. Table 3 lists the results of this experiment.

Table 3: Accuracy of 'hot topic words' (Strict Standard of Assessment)

	all	non-rss	rss	random	rss-desc
May	60%	67%	53%	63%	50%
Jun.	57%	57%	37%	47%	43%
Jul.	80%	87%	70%	80%	90%
Aug.	80%	77%	77%	77%	63%
Avr.	69%	72%	59%	67%	62%

This table shows that dataset 'rss' and dataset 'rss-desc' have worse accuracy than the others. They are significantly worse than 'non-rss' (using chi-square test at 5% level). However, it is true that the results were more influenced by the evaluator's subjectivity than in the previous experiment.

5.3 Number of Topics 'Hot topic words' Contained

In this section, we discuss our investigation into the number of topics that the 'hot topic words' extracted from each dataset contain. A topic was assigned to each 'hot topic word' by the evaluator, and the number of unique topics assigned to each dataset per month was counted. The number of unique topics indicates how many topics each dataset contains, and this means that how the writers' interests in the topics is distributed.

'Hot topic words' were assessed to be the same topic when two words were completely the same, synonyms, parts of a collocation, or misanalyzed with a morphological analyzer. We also judged all words that represented the names of racehorses to belong to the same topic.

Table 4 lists the result of this experiment.

Table 4: Number of topics which 'hot topic words' contain

	all	non-rss	rss	random	rss-desc
May	15	16	14	14	12
Jun.	14	16	10	14	12
Jul.	14	14	12	13	12
Aug.	14	14	13	13	17
Avr.	14.25	15.00	12.25	13.50	13.25

The numbers in the table correspond to the number of topics that each dataset contained for a month. As you can see, 'non-rss' contains more topics than the others. The reason can be considered to be that 'non-rss' contains much more variety of topics than the other datasets. However, the differences are not statistically significant.

5.4 Categories 'Hot topic words' Belong to

In the previous section, we assigned a topic to each 'hot topic word'. We then classified these topics into categories. This experiment revealed which category each dataset tended to contain.

We used seven categories of 'Sports', 'Animations/Games/PCs', 'Politics/Economics/Incidents', 'Entertainments/Culture', 'Weather', 'Holidays/Seasonal Events' and 'Science'.

Table 5 lists the number of 'hot topic words' belonging to these categories.

Table 5: Categories 'hot topic words' belong to

	all	non-rss	rss	random	rss-desc	total†
Sports	53	31	64	50	46	94
Animations	4	8	1	4	1	9
Politics	16	20	23	18	11	32
Entertainments	12	11	11	9	5	22
Weather	11	11	9	10	8	16
Holidays	21	24	34	23	21	33
Science	0	0	1	2	0	2

† total number of distinct 'hot topic words' which belong the category

The numbers in the table represent the total number of words for months, and the column 'variety' shows the total number of unique topics for each category. The number of 'sports' for 'rss' is larger than the number for 'non-rss', and conversely the number of 'Animations/Games/PCs' for 'non-rss' are larger than 'rss'.

5.5 Agreement between Topics 'Hot topic words' Represent

Last, we used the topics assigned to the words in Section 5.3, and investigated the degree of agreement of the topics in two datasets. If the degree of agreement between two datasets is high, we can consider that these datasets have similar content.

The results are listed in Table 6. The numbers in the table indicate the degree of overlaps where the topics in the row dataset agree with the ones in the column dataset. For example, 0.62 at the top right of the table means 62% of 'hot topic words' for 'all' agree with 'rss-desc'.

Table 6: Agreement between topics 'hot topic words' represent

	all	non-rss	rss	random	rss-desc
all	-	0.88	0.73	0.82	0.62
non-rss	0.83	-	0.59	0.75	0.60
rss	0.71	0.63	-	0.77	0.59
random	0.79	0.77	0.76	-	0.59
rss-desc	0.58	0.59	0.59	0.59	-

As you can see, 'rss-desc' has the low degree of agreement with the others, and the degree of agreement between 'non-rss' and 'rss' is also low.

6. DISCUSSION AND CONCLUSION

The experiments in the previous section revealed the following.

First, none of the datasets had no serious deviations and could appropriately be used to extract hot topics with the burst identification algorithm since the results of the first experiment were highly precise. However, when we used

stricter standard of assessment, the results from the datasets of blogs ('rss' and 'rss-desc') were worse than those from the 'web diaries' ('non-rss'). 'Non-rss' contained more topics than the other datasets. This suggests that 'web diaries' contained more distributed topics than blogs. However, this was not statistically significant. We also investigated the hot topic categories for each dataset. The results showed that blogs contained many topics related to 'sports', and 'web diaries' contained many topics related to 'Animations/Games/PCs'. The degree of agreement between blogs ('rss') and 'web diaries' ('non-rss') was low. This also suggests the writers/readers of blogs and 'web diaries' had different interests. We therefore conluded that there was a difference between blog and 'web diary' content.

Considering the results from the point of view of text mining, 'rss-desc' had significantly worse precision, as discussed in Section 5.1. This was caused by the short text given by the description element of RSS. We could obtain the better results by using not only RSS metadata but also the full text of the entries on text mining. The experiments described in this paper targeted 2003 and 2004 when there were adequate blog entries. However, as there were not that many blog entries published before then, entries of 'web diaries' for these years would have been a big advantage for text mining.

However, not much data was analyzed, and only one person evaluated the data. A way to resolve this problem would be to increase the number of analysts, which we intend to do in the future.

We investigated the differences between blogs and 'web diaries' by examining 'hot topic words'; however, there are still some other approaches remaining to pursue. For example, these are to compare the out-links/in-links of blogs and 'web diaries', to investigate the differences in average entry length or modality of sentences, and to conduct follow-up research of blog users who previously wrote 'web diaries'. These remain for future work.

Acknowledgments

This work was supported by The 21st Century COE Program "Framework for Systematization and Application of Large-scale Knowledge Resources", Japan Society for the Promotion of Science.

7. REFERENCES

[1] Rdf rich site summary (rss). `http://www.oasis-open.org/cover/rss.html`.

[2] T. Fujiki, T. Nanno, Y. Suzuki, and M. Okumura. Identification of bursts in a document stream. In *First International Workshop on Knowledge Discovery in Data Streams (in conjunction with ECML/PKDD 2004)*, 2004.

[3] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1–25, 2002.

[4] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of the 12th International World Wide Web Conference*, pages 568–576, 2003.

[5] A. Miura and K. Yamashita. Why do people publish weblogs?: An online survey of weblog authors in japan. In *Human Perspectives in the Internet Society: Culture, Psychology and Gender*, 2004.

[6] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining japanese weblogs. In *Poster Session of the 13th International World Wide Web Conference*, 2004.

[7] T. Nanno, Y. Suzuki, T. Fujiki, and M. Okumura. Automatic collection and monitoring of japanese weblogs. In *WWW2004 1st Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics of the 13th International World Wide Web Conference*, 2004.

[8] D. Winer. Weblogs.com xml-rpc interface. `http://www.xmlrpc.com/weblogsCom`, 2001.

[9] K. Yamashita and A. Miura. Why do people keep writing web diaries and weblogs? (in japanese). In *45th Annual Meeting of the Japanese Psychological Association*, 2004.