Extracting Topics From Weblogs Through Frequency Segments

Mizuki Oka
University of Tsukuba, Japan
mizuki@osss.cs.tsukuba.ac.jp

Hirotake Abe
CREST, Japan Science and Technology Agency
habe@osss.cs.tsukuba.ac.jp

Kazuhiko Kato
University of Tsukuba, Japan
kato@cs.tsukuba.ac.jp

ABSTRACT

In this paper, we present an approach to extracting topics from weblogs by using terms that appear in them. We model a term in terms of *frequency segments*, i.e., sequential occurrences of the term over time, as the unit of characterization. A notable feature of the model is its approximation of changes in the dynamics of term frequencies; it captures the granularity of frequencies from the very beginning of their occurrence. This approximation also makes a comparison of frequency patterns of terms more effective. We report on the results obtained from weblogs that contained an event of global significance i.e., the London bombings of 2005.

Keywords

Topic characterization, topic detection, blogs

1. INTRODUCTION

Weblogs have become a key tool not only for individuals to publish posts, but also for obtaining useful information on a daily basis. By providing tools that make publishing easy, blog sites have made information publishing significantly more efficient. The number of articles posted each day to sites indicates their popularity. For example, large blog sites such as *Boing Boing* [1], *Engadget* [2], and *Google Weblog* [3] publish tens of millions of posts per day. Weblogs can obviously be used to identify trends and important events or subjects if we carefully analyzed the posts.

Tracking topics and their descriptions over a period of time can serve to describe changes in trends. This information could then be used, for example, by blog search services to provide users with a series of terms related to a given input query in the form of a timeline or to predict future changes in the trends.

One common approach to analyzing weblogs is to use hyperlinks and their structure. For example, Kumar et al. [9] used the evolving link structure to capture bursts of activity within blog communities. Another approach is to use the text of weblogs. Gruhl et al. [8] used the text to characterize information diffusion by capturing the propagation of topics from one blog to the next. Like Gruhl et al., we used the text of weblogs to mine terms that described topics. That is, we first sought to identify a collection of main terms that were characteristic to some topics. We propose that a term can be characterized and ranked using the concept of *frequency segments*, or simply *segments* (sequential occurrences of a term over time), i.e., the parameters derived from segments effectively capture the dynamics of changes in the occurrences of a term over a period of time. We found the relations between identified terms that describe a topic to track it by the flow of terms over time.

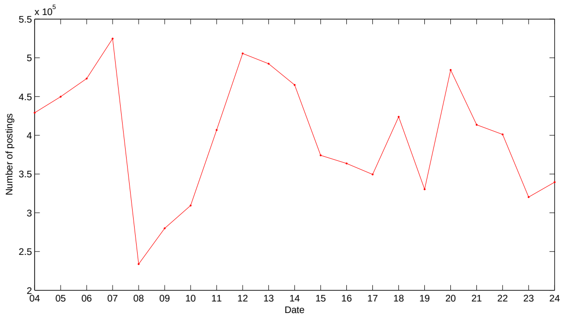

Figure 1: Number of blog postings per day

We used the blog corpus provided by Intelliseek [4] at the Third Annual Workshop on the Weblogging Ecosystem [5]. The corpus included annotations such as the dates of posting, the times of posting, author's names, the titles of the posts, weblog URLs, permalinks, tags/categories, and outlinks classified by type. The data were of the period from July 4 through July 24, 2005, a period in which an event of global significance occurred, i.e., the London bombings. They contained a total of 8,370,193 postings. The number of weblog postings per day is plotted in Fig.1.

The rest of the paper is organized as follows. In Section 2, we present our approach to characterizing and ranking terms from a series of weblogs with examples from our experimental results. In Section 3, we discuss related work. We end with a summary and an outline of our future research direction in Section 4.

2. MODELING OF TOPICS

We explore the *topics* discussed in the blog corpus in this section. For our preliminary experiments, we used 2000 weblogs each day, accounting for a total of 42,000 postings in total (0.5% percent of all postings in the corpus). We then extracted only nouns from the postings and used them as

WWW2006, May 22–26, 2006, Edinburgh, UK.
.

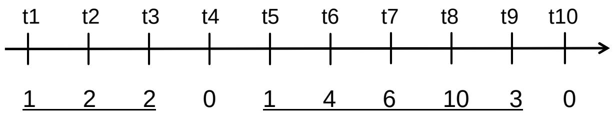

Figure 2: Sequence of frequency of occurrences of a term in each unit of time

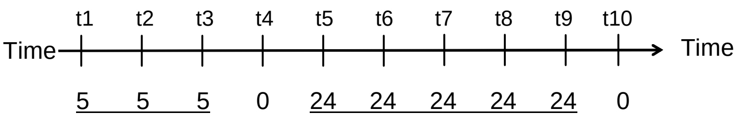

Figure 3: Segment sum sequence for [1 2 2 0 1 4 6 10 3 0]

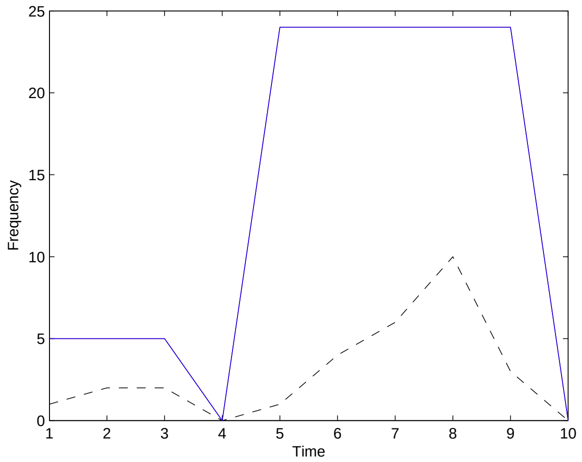

Figure 4: Frequency of occurrence and segment sum for [1 2 2 0 1 4 6 10 3 0]

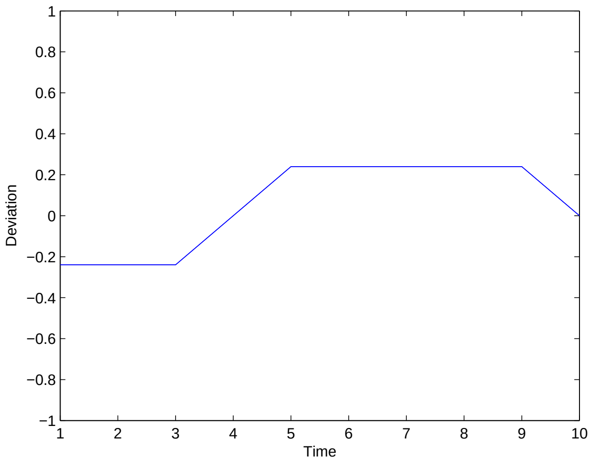

Figure 5: Deviations of segment sum for [1 2 2 0 1 4 6 10 3 0].

terms in our analysis.

2.1 Characterization and Identification of Terms

Identifying *important* term is the first step in finding topics. This problem of identification has been studied in depth in the field of topic detection and tracking for a number of years [7]. The objectives of previous research have been to find event-based topics from news media and/or radio with a high degree of precision or recall. Our objectives were somewhat different; we required methods that would extract comprehensive subtopics that were structurally related to a topic regardless of its granularity. We thus propose characterizing terms using the concept of successive occurrences of terms, which we called *frequency segments*. To capture the dynamics of changes in the occurrences of a term over a period of time, we use parameters such as *segment sum* and *segment deviation*. We then rank terms to identify a set of *important* terms that describe topics. Finally, we relate terms by finding terms that occur together in the same postings. We then track a topic by connecting the related terms over time.

2.1.1 Frequency Segments

Given sequence V of T elements, we measure the dynamics of V in units of segments. Here, a segment is a block of sequential occurrences of a term. For example, given a sequence of occurrences $(1, 2, 2, 0, 1, 4, 6, 10, 30)$, segments are identified as $(1, 2, 2)$ and $(1, 4, 6, 10, 3)$ (see Fig. 2).

We characterize terms using segments with two parameters corresponding to the *absolute* strength in the weight of occurrences (total sum of occurrences of each segment, which we call segment sum and denote it as $sum_{segment}$) and the *relative* strength with respect to the average strength in the weight of occurrences (deviation in the total sum of occurrences of each segment, which we call segment deviation and denote it as $dev_{segment}$). Segment deviation is defined as:

$$\begin{aligned} z &= \frac{sum_{segment} - mean(sum_{segment})}{3\sigma} \\ dev_{segment} &= \begin{cases} min(z, 1.0), & \text{if } z \geq 0 \\ max(-1.0, z), & \text{othersize,} \end{cases} \end{aligned}$$

where σ indicates the standard deviation. In the example in Fig. 2, we have $sum_{(1,2,2)} = 5$, $dev_{(1,2,2)} = -0.24$, $sum_{(1,4,6,10,3)} = 24$, and $dev_{(1,4,6,10,3)} = 0.24$. We assign the same values for sum and deviation to the same segment. For example, the segment sum in the example sequence is depicted in Fig. 3. This replacement of frequency of occurrences with segment sums enhances the dynamics of changes in the frequency of occurrences over time. In other words, the granularity of frequencies are already captured from the beginning of occurrences. Figure 4 plots the frequency of occurrences and their segment sums in the example sequence. We then captured the enhanced dynamics of changes in the segment sums in terms of their deviations. Figure 5 plots deviations in the example sequence.

There is a plot of the deviations and sums of segments for all the terms in the analyzed corpus in Fig. 6. A notable feature in the figure is the vertical line along the zero value on the x-axis (sum deviation). A term with a zero value for sum deviations for the entire time period indicates that

information	news	research	Web
company	food	blog	city
money	service	marketing	car
system	world	business	industry
wedding	life	book	movie
music	home	job	history

Table 1: Sample of terms whose segment deviations equal zero

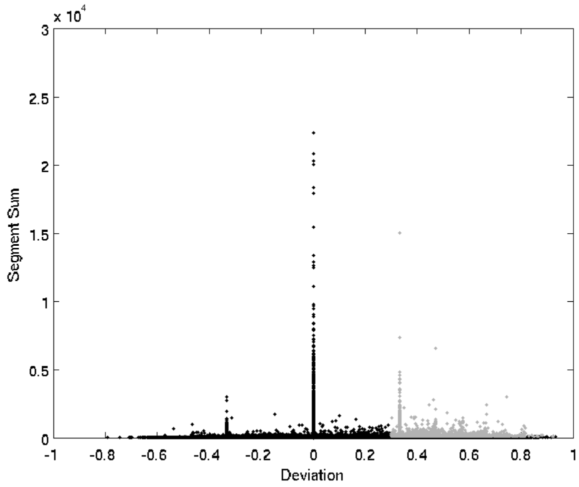

Figure 6: Deviations and sums of segments of all terms in 41000 postings.

there is only one segment in the sequence of occurrences of the term, indicating that the term occurs everyday. The total number of such terms was about 11,000. We analyzed terms that occurred more than 3000 times per day (total of about 150 terms) and found that most of them were general; samples of these are listed in Table 1.

To obtain "interesting" terms, we hypothesized that they would have relatively high values for both deviations and sums because they represented greater and/or sudden increases in occurrences. We thus generated a collection of relevant terms using a threshold of $dev_{segment} > 0.3$ and $sum_{segment} > 90$, accounting for about 3,200 terms, which appear in light gray in Fig. 6.

2.1.2 Ranking Terms and Finding Related Terms

Having identified the set of terms to consider, our next goals were to extract the main terms for each day in the analyzed corpus and to find related terms to track a topic. To extract the main terms, we scored term t to rank the terms with the function :

$$score(t) = \alpha \times log(sum_{segment}) + dev_{segment},$$

where α is determined empirically (in this experiment, we used $\alpha = 0.11$). We then related terms that co-occurred in the same posting.

The top seven ranked terms for the period July 4 through 13, including the London Bombing attack on the 7th are listed in Table 2. As we can observe from the table, terms such as **G8**, **Summit**, and **underground**, which represent events of significance in July 2005, are highly ranked. We analyzed terms related to **G8** in the top seven ranked terms and marked them in bold with a larger font in the table. It is interesting to note that terms such as **Carey** and **Angelina** appeared in the top seven ranked terms and were identified as being related to G8 (Carey indicates the famous singer Mariah Carey). An examination into the relation between Carey and G8 first revealed that they appeared as related because of the concert called Live 8 designed to coincide with the G8 summit in which Mariah Carey was a participant. The term Angelina applies to an American actress called Angelina Jolie. An examination into her relation with G8 revealed that the term appeared first because she adopted a child in the same time period as G8 and second this adoption raised rumors about her relationship with the American actor Brad Pitt, who gave a talk at the same concert.

Rank	Term	Rank	Term
1	class	11	blair
2	tube	12	G8
3	terrorism	13	rail
4	tragedy	14	jessica
5	bomber	15	contest
6	bombing	16	truck
7	unlimited	17	tournament
8	evening	18	preparation
9	suite	19	aruba
10	minister	20	democratic

Table 3: Terms related to underground

2.2 Extraction of Detailed Description of Topics

Although tracking the identified main terms discussed above makes it possible to capture an overview of topics, they do not offer more detailed descriptions of each topic. Thus in this section, we consider the dynamic structure of terms to extract such detailed terms related to a topic using segment deviation patterns. As an example for analysis, we selected the term *underground* that was highly ranked in the previous analysis and analyzed its related terms. Terms that co-occured with *underground* accounted for a total of about 320 terms. We were interested in extracting terms from these related terms that had similar segment deviation patterns with the target term, in this case *underground.*

We analyzed the segment deviation patterns of terms related to underground from July 6 through 12, the period in which the term underground was identified as being important for about one day. Figure 7 plots the segment deviation pattern for the term *underground.* Figure 8 is for the term bombing and Fig. 9 is for terrorism. We can see that they have very similar deviation patterns with the term underground for this period of time, whereas the term *judge* in Fig. 10 which was also identified as related to the term underground, has no similarity in the deviation pattern.

Using the concept of segments to characterize terms ensures the granularity of a term at its starting point, and comparing their patterns becomes fairly simple. We thus used the simple function to compute the similarity in deviation patterns of related terms. That is, given deviation

Date/Rank	1	2	3	4	5	6	7
Jul. 4	GM	**G8**	skull	permalink	Remedies	paint	**Carey**
Jul. 5	**G8**	Candida	permalink	Remedies	paint	**Carey**	payday
Jul. 6	**G8**	Candida	bridal	Remedies	paint	payday	**summit**
Jul. 7	vinyl	**G8**	dividend	**underground**	bridal	nba	valve
Jul. 8	vinyl	**G8**	**underground**	bridal	nba	valve	paint
Jul. 9	**G8**	**underground**	valve	**Democratic**	Jo	Wars	Career
Jul.10	Lcd	**G8**	lah	otra	**underground**	Kodak	dan
Jul. 11	**G8**	lah	Simpson	**underground**	dan	Bath	Accounting
Jul. 12	Simpson	dan	carpet	Bath	Cooper	haha	scanner
Jul. 13	Simpson	dan	carpet	Bath	haha	Manual	**Angelina**

Table 2: Top seven ranked terms for each day and terms related to G8

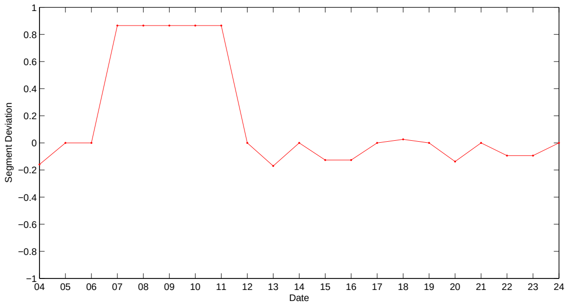

Figure 7: Segment deviation for *underground*

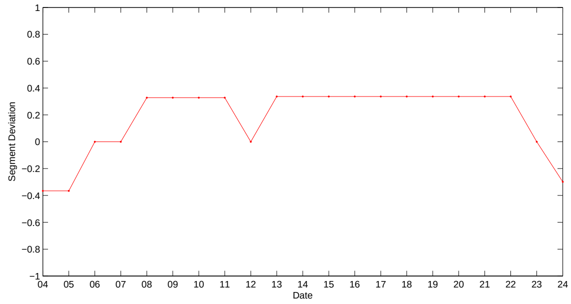

Figure 8: Segment deviation for *bombing*

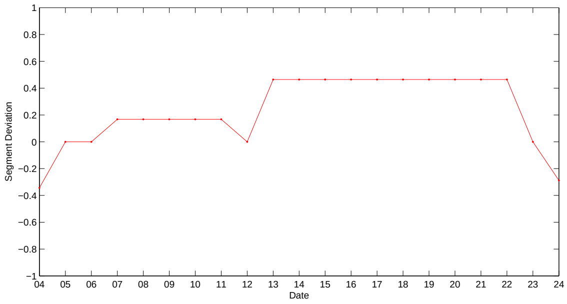

Figure 9: Segment deviation for *terrorism*

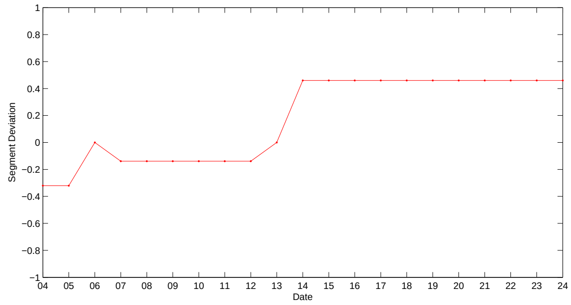

Figure 10: Segment deviation for *judge*

vectors of $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_n)$, the distance between A and B are computed by

$$correlation(A, B) = \frac{< A - mean(A), B - mean(B) >}{\|(A - mean(A))\| \, \|(B - mean(B))\|},$$

where $mean(X)$ denotes the mean value of X, $< X, Y >$ denotes the inner product of X and Y, and $\|X\|$ denotes the norm of X. Table 3 lists the top 20 ranked terms related to the term *underground* using this function. We can see that terms such as *tube*, *terrorism*, *tragedy*, *bomber*, and *bombing* are highly ranked, which coincide with human perception.

3. RELATED WORK

Related work on the detection of topics from weblogs exist either as research goals themselves and/or one of the steps towards the detection of communities, trends, important bloggers, and flow of information. Kumar et al. [9] studied the extraction of bursty communities from Blogspace through temporal bursts of hyperlinkings using postings as units for their analysis. Adar et al. [6] studied the implicit structure and dynamics of blogspace examining both its macro and micro behaviors. Nakajima et al. [10] worked on discovering important bloggers considering weblogs as conversations (thread) and analyzed the patterns for the income and outcome weblogs of bloggers.

While the above researchers used links and their structures to analyze weblogs, Gruhl et al. [8] focused on the propagation of topics from one blog to the next, based on the *text* of weblogs as we have done. In their work, they found interesting terms (e.g., nouns) by ranking term t by the ratio of the number of times that t is mentioned on a particular day, i to the average number of times t was mentioned on previous days, which they called the $tfcidf$ score. They then characterized the topic structure in terms of *chatter* and *spikes* quantifying them using two parameters, i.e., the distribution for the number of posts per day (chatter level) and the distributions for the frequency, volume, and shapes of spikes (spike pattern). Their work was similar to ours to the extent that they extracted interesting terms by ranking terms and quantifying them. However, our approach is based on the concept of frequency segments over several days instead of the frequency of occurrences of terms per day. Because the frequency segments ensure the granularity of a term at its starting point, topic can be detected more effectively.

4. CONCLUSION AND FUTURE WORK

We presented an approach to extracting topics from weblogs by charactering terms using the sequential occurrences of terms over time, which we referred to as frequency segments in this paper. A neat feature of the model we presented is its approximation of changes in the dynamics of term frequencies, i.e., it captures the granularity of frequencies from the very beginning of their occurrence. Using frequency segments, we characterized terms with two parameters corresponding to the absolute strength, in the weight of occurrences (segment sum), and the relative strength, with respect to the average strength in the weight of occurrences (segment deviation). Experiments were conducted on the corpus of weblogs that contained events on the London Bombings of July 2005. As a result, we could successfully detect a set of terms that described this event by ranking terms using the two parameters. We also analyzed related to capture more descriptive terms to obtain more detailed information on the events and we obtained terms that coincided with human perception.

In this paper, the focused on detecting terms that described topics, but seeking to predict the future dynamics of topics would be interesting. Since our model makes a comparison of frequency patterns more effective by approximation, this could be done by analyzing the dynamics in their changes more closely.

5. ACKNOWLEDGMENTS

We wish to thank Professor Manabu Okumura, Dr. Tomoyuki Nanno, and Mr. Toshiaki Fujiki of the Tokyo Institute of Technology for their stimulating discussions and useful suggestions to support our research.

This work was partly supported by a Grant-in-Aid for JSPS Fellows, made available by the Japan Society for the Promotion of Science.

6. REFERENCES

[1] Boing boing http://www.mysql.com/.

[2] Engadget http://www.engadget.com/.

[3] Google weblog http://google.blogspace.com/.

[4] Intelliseek http://www.invisibleweb.com/.

[5] Third annual workshop on the weblogging ecosystem http://www.blogpulse.com/www2006-workshop/.

[6] E. Adar and L. Zhang. Implicit structure and the dynamics of blogspace. In *WWW '04 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[7] J. Allan. *Topic Detection and Tracking: Event-Based Information Organization*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[8] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW '04: Proceedings of the 13th International Conference on World Wide Web*, pages 491–501, New York, NY, USA, 2004. ACM Press.

[9] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th International Conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[10] S. Nakajima, J. Tatemura, Y. Hino, Y. Hara, and K. Tanaka. Discovering important bloggers based on analyzing blog threads. In *WWW '05 Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2005.