Event Detection and Summarization in Weblogs with Temporal Collocations

Chun-Yuan Teng and Hsin-Hsi Chen

Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
{r93019, hhchen}@csie.ntu.edu.tw

Abstract

This paper deals with the relationship between weblog content and time. With the proposed temporal mutual information, we analyze the collocations in time dimension, and the interesting collocations related to special events. The temporal mutual information is employed to observe the strength of term-to-term associations over time. An event detection algorithm identifies the collocations that may cause an event in a specific timestamp. An event summarization algorithm retrieves a set of collocations which describe an event. We compare our approach with the approach without considering the time interval. The experimental results demonstrate that the temporal collocations capture the real world semantics and real world events over time.

1. Introduction

Compared with traditional media such as online news and enterprise websites, weblogs have several unique characteristics, e.g., containing abundant life experiences and public opinions toward different topics, highly sensitive to the events occurring in the real world, and associated with the personal information of bloggers. Some works have been proposed to leverage these characteristics, e.g., the study of the relationship between the content and bloggers' profiles (Adamic & Glance, 2005; Burger & Henderson, 2006; Teng & Chen, 2006), and content and real events (Glance, Hurst & Tornkiyo, 2004; Kim, 2005; Thelwall, 2006; Thompson, 2003).

In this paper, we will use *temporal collocation* to model the term-to-term association over time. In the past, some useful collocation models (Manning & Schütze, 1999) have been proposed such as mean and variance, hypothesis test, mutual information, etc. Some works analyze the weblogs from the aspect of time like the dynamics of weblogs in time and location (Mei, et al., 2006), the weblog posting behavior (Doran, Griffith & Henderson, 2006; Hurst, 2006), the topic extraction (Oka, Abe & Kato, 2006), etc. The impacts of events on social media are also discussed, e.g., the change of weblogs after London attack (Thelwall, 2006), the relationship between the warblog and weblogs (Kim, 2005; Thompson, 2003), etc.

This paper is organized as follows. Section 2 defines *temporal collocation* to model the strength of term-to-term associations over time. Section 3 introduces an *event detection algorithm* to detect the events in weblogs, and an *event summarization algorithm* to extract the description of an event in a specific time with temporal collocations. Section 4 shows and discusses the experimental results. Section 5 concludes the remarks.

2. Temporal Collocations

We derive the temporal collocations from Shannon's mutual information (Manning & Schütze, 1999) which is defined as follows (Definition 1).

Definition 1 (Mutual Information) The mutual information of two terms *x* and *y* is defined as:

$$I(x,y) = P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

where $P(x,y)$ is the co-occurrence probability of *x* and *y*, and $P(x)$ and $P(y)$ denote the occurrence probability of *x* and *y*, respectively.

Following the definition of mutual information, we derive the temporal mutual information modeling the term-to-term association over time, and the definition is given as follows.

Definition 2 (Temporal Mutual Information) Given a timestamp *t* and a pair of terms *x* and *y*, the temporal mutual information of *x* and *y* in *t* is defined as:

$$I(x,y\mid t) = P(x,y\mid t)\log\frac{P(x,y\mid t)}{P(x\mid t)P(y\mid t)}$$

where $P(x,y|t)$ is the probability of co-occurrence of terms *x* and *y* in timestamp *t*, $P(x|t)$ and $P(y|t)$ denote the probability of occurrences of *x* and *y* in timestamp *t*, respectively.

To measure the change of mutual information in time dimension, we define the change of temporal mutual information as follows.

Definition 3 (Change of Temporal Mutual Information) Given time interval $[t_1, t_2]$, the change of temporal mutual information is defined as:

$$C(x,y,t_1,t_2) = \frac{I(x,y\mid t_2) - I(x,y\mid t_1)}{t_2 - t_1}$$

where $C(x,y,t_1,t_2)$ is the change of temporal mutual information of terms *x* and *y* in time interval $[t_1, t_2]$, $I(x,y|t_1)$ and $I(x,y|t_2)$ are the temporal mutual information in time t_1 and t_2, respectively.

3. Event Detection

Event detection aims to identify the collocations resulting in events and then retrieve the description of events. Figure 1 sketches an example of event detection. The weblog is parsed into a set of collocations. All collocations are processed and monitored to identify the plausible events. Here, a regular event "Mother's day" and an irregular event "Typhoon Chanchu" are detected. The event "Typhoon Chanchu" is described by the words

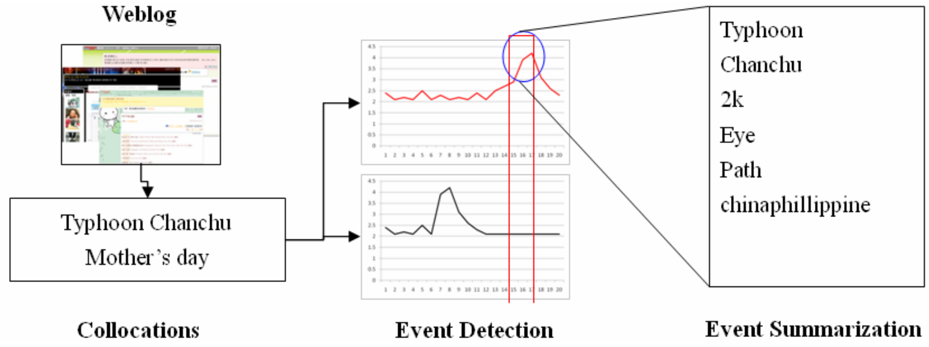


Figure 1: An Example of Event Detection

“Typhoon”, “Chanchu”, “2k”, “Eye”, “Path” and “chinaphillippine”.

The architecture of an event detection system includes a preprocessing phase for parsing the weblogs and retrieving the collocations; an event detection phase detecting the unusual peak of the change of temporal mutual information and identifying the set of collocations which may result in an event in a specific time duration; and an event summarization phase extracting the collocations related to the seed collocations found in a specific time duration.

The most important part in the preprocessing phase is collocation extraction. We retrieve the collocations from the sentences in blog posts. The candidates are two terms within a window size. Due to the size of candidates, we have to identify the set of tracking terms for further analysis. In this paper, those candidates containing stopwords or with low change of temporal mutual information are removed.

In the event detection phase, we detect events by using the peak of temporal mutual information in time dimension. However, the regular pattern of temporal mutual information may cause problems to our detection. Therefore, we remove the regular pattern by seasonal index, and then detect the plausible events by measuring the unusual peak of temporal mutual information.

If a topic is suddenly discussed, the relationship between the related terms will become higher. Two alternatives including change of temporal mutual information and relative change of temporal mutual information are employed to detect unusual events. Given timestamps t_1 and t_2 with temporal mutual information MI_1 and MI_2, the change of temporal mutual information is calculated by $(MI_2\text{-}MI_1)$. The relative change of temporal mutual information is calculated by $(MI_2\text{-}MI_1)/MI_1$.

For each plausible event, there is a seed collocation, e.g., “Typhoon Chanchu”. In the event description retrieval phase, we try to select the collocations with the highest mutual information with the word *w* in a seed collocation. They will form a collocation network for the event. Initially, the seed collocation is placed into the network. When a new collocation is added, we compute the mutual information of the multiword collocations by the following formula, where *n* is the number of collocations in the network up to now.

$$\mathit{Multiword\ Mutual\ Information} = \prod_{i=1}^{n} MI_i$$

If the multiword mutual information is lower than a threshold, the algorithm stops and returns the words in the collocation network as a description of the event. Figure 2 sketches an example. The collocations “Chanchu’s path”, “Typhoon eye”, and “Chanchu affects” are added into the network in sequence based on their MI.

We have two alternatives to add the collocations to the event description. The first method adds the collocations which have the highest mutual information as discussed above. In contrast, the second method adds the collocations which have the highest product of mutual information and change of temporal mutual information.

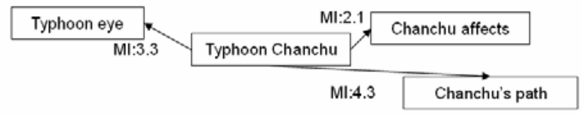


Figure 2: An Example of Collocation network

4. Experiments and Discussions

4.1. Temporal Mutual Information versus Mutual Information

In the experiments, we adopt the ICWSM weblog data set (Teng & Chen, 2007; ICWSM, 2007). This data set collected from May 1, 2006 through May 20, 2006 is about 20 GB. Without loss of generality, we use the English weblog of 2,734,518 articles for analysis.

To evaluate the effectiveness of time information, we made the experiments based on mutual information (Definition 1) and temporal mutual information (Definition 2). The former called the *incremental approach* measures the mutual information at each time point based on all available temporal information at that time. The latter called the *interval-based approach* considers the temporal mutual information in different time stamps. Figures 3 and 4 show the comparisons between interval-based approach and incremental approach, respectively, in the event of Da Vinci Code.

We find that “Tom Hanks” has higher change of temporal mutual information compared to “Da Vinci Code”. Compared to the incremental approach in Figure 4, the interval-based approach can reflect the exact release date of “Da Vinci Code.”

4.2. Evaluation of Event Detection

We consider the events of May 2006 listed in wikipedia[1] as gold standard. On the one hand, the events posted in wikipedia are not always complete, so that we adopt recall rate as our evaluation metric. On the other hand, the events specified in wikipedia are not always discussed in weblogs. Thus, we search the contents of blog post to verify if the events were touched on in our blog corpus. Before evaluation, we remove the events listed in wikipedia, but not referenced in the weblogs.

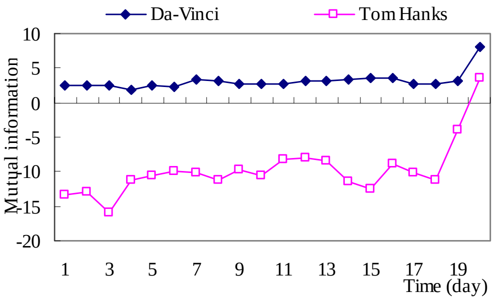


Figure 3: Interval-based Approach in Da Vinci Code

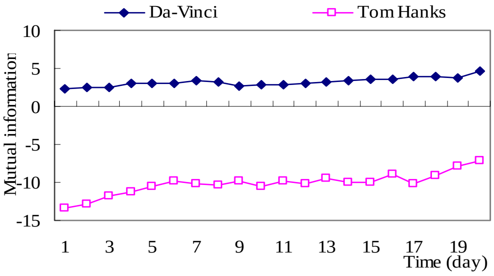


Figure 4: Incremental Approach in Da Vinci Code

Figure 5 sketches the idea of evaluation. The left side of this figure shows the collocations detected by our event detection system, and the right side shows the events listed in wikipedia. After matching these two lists, we can find that the first three listed events were correctly identified by our system. Only the event "Nepal Civil War" was listed, but not found. Thus, the recall rate is 75% in this case.

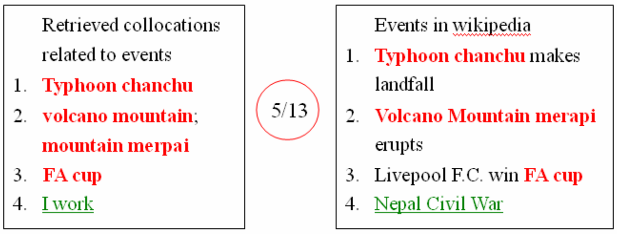

Retrieved collocations related to events
1. Typhoon chanchu
2. volcano mountain; mountain merpai
3. FA cup
4. I work
5/13
Events in wikipedia
1. Typhoon chanchu makes landfall
2. Volcano Mountain merapi erupts
3. Livepool F.C. win FA cup
4. Nepal Civil War

Figure 5: Evaluation of Event Detection Phase

As discussed in Section 3, we adopt change of temporal mutual information, and relative change of temporal mutual information to detect the peak. In Figure 6, we compare the two methods to detect the events in weblogs. The relative change of temporal mutual information achieves better performance than the change of temporal mutual information.

Table 1 and Table 2 list the top 20 collocations based on these two approaches, respectively. The results of the first approach show that some collocations are related to the feelings such as "fell left" and time such as "Saturday night". In contrast, the results of the second approach show more interesting collocations related to the news events at that time, such as terrorists "zacarias moussaoui" and "paramod mahajan." These two persons were killed in May 3. Besides, "Geena Davis" got the golden award in May 3. That explains why the collocations detected by relative change of temporal mutual information are better than those detected by change of temporal mutual information.

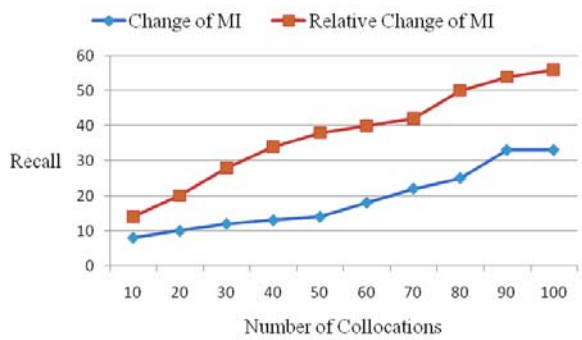


Figure 6: Performance of Event Detection Phase

Collocations	CMI	Collocations	CMI
May 03	9276.08	Current music	1842.67
Illegal immigrants	5833.17	Hate studying	1722.32
Feel left	5411.57	Stephen Colbert	1709.59
Saturday night	4155.29	Thursday night	1678.78
Past weekend	2405.32	Can't believe	1533.33
White house	2208.89	Feel asleep	1428.18
Red sox	2208.43	Ice cream	1373.23
Album tool	2120.30	Oh god	1369.52
Sunday morning	2006.78	Illegalimmigration	1368.12
Sunday night	1992.37	Pretty cool	1316.56

Table 1: Top 20 collocations with highest change of temporal mutual information

Collocations	CMI	Collocations	CMI
casinos online	618.36	Diet sodas	32.50
zacarias moussaoui	154.68	Ving rhames	31.63
Tsunami warning	107.93	Stock picks	29.09
Conspirator zacarias	71.62	Happy hump	28.45
Artist formerly	57.04	Wong kan	28.34
Federal Jury	41.78	Sixapartcom movabletype	28.13
Wed 3	39.20	Aaron echolls	27.48
Pramod mahajan	35.41	Phnom penh	25.78
BBC Version	35.21	Livejournal sixapartcom	23.83
Geena davis	33.64	George yeo	20.34

Table 2: Top 20 collocations with highest relative change of mutual information

4.3. Evaluation of Event Summarization

As discussed in Section 3, we have two methods to include collocations to the event description. Method 1 employs the highest mutual information, and Method 2

[1] http://en.wikipedia.org/wiki/May_2006

utilizes the highest product of mutual information and change of temporal mutual information. Figure 7 shows the performance of Method 1 and Method 2. We can see that the performance of Method 2 is better than that of Method 1 in most cases.

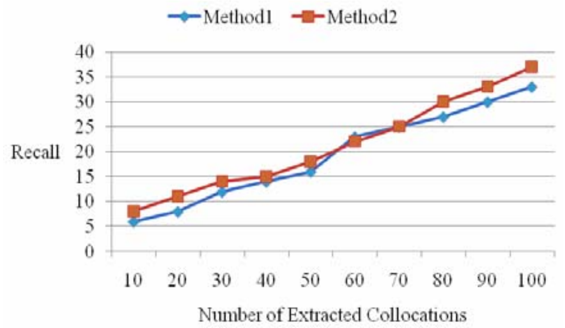
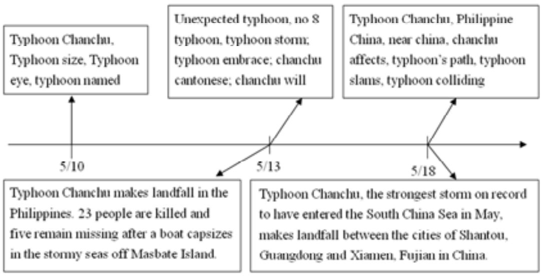


Figure 7: Overall Performance of Event Summarization

The results of event summarization by Method 2 are shown in Figure 8. Typhoon Chanchu appeared in the Pacific Ocean on May 10, 2006, passed through Philippine and China and resulted in disasters in these areas on May 13 and 18, 2006. The appearance of the typhoon Chanchu cannot be found from the events listed in wikipedia on May 10. However, we can identify the appearance of typhoon Chanchu from the description of the typhoon appearance such as “typhoon named” and “Typhoon eye. In addition, the typhoon Chanchu’s path can also be inferred from the retrieved collocations such as “Philippine China” and “near China”. The response of bloggers such as “unexpected typhoon” and “8 typhoons” is also extracted.


Typhoon Chanchu, Typhoon size, Typhoon eye, typhoon named
Unexpected typhoon, no 8 typhoon, typhoon storm; typhoon embrace; chanchu cantonese; chanchu will
Typhoon Chanchu, Philippine China, near china, chanchu affects, typhoon's path, typhoon slams, typhoon colliding
5/10
5/13
5/18

Typhoon Chanchu makes landfall in the Philippines. 23 people are killed and five remain missing after a boat capsizes in the stormy seas off Masbate Island.

Typhoon Chanchu, the strongest storm on record to have entered the South China Sea in May, makes landfall between the cities of Shantou, Guangdong and Xiamen, Fujian in China.

Figure 8: Event Summarization for Typhoon Chanchu

5. Concluding Remarks

This paper introduces temporal mutual information to capture term-term association over time in weblogs. The extracted collocation with unusual peak which is in terms of relative change of temporal mutual information is selected to represent an event. We collect those collocations with the highest product of mutual information and change of temporal mutual information to summarize the specific event. The experiments on ICWSM weblog data set and evaluation with wikipedia event lists at the same period as weblogs demonstrate the feasibility of the proposed temporal collocation model and event detection algorithms.

Currently, we do not consider user groups and locations. This methodology will be extended to model the collocations over time and location, and the relationship between the user-preferred usage of collocations and the profile of users.

Acknowledgments

Research of this paper was partially supported by National Science Council, Taiwan (NSC96-2628-E-002-240-MY3) and Excellent Research Projects of National Taiwan University (96R0062-AE00-02).

References

Adamic, L.A., Glance, N. (2005). The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. In: *Proceedings of the 3rd International Workshop on Link Discovery*, pp. 36--43.

Burger, J.D., Henderson J.C. (2006). An Exploration of Observable Features Related to Blogger Age. In: *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 15--20.

Doran, C., Griffith, J., Henderson, J. (2006). Highlights from 12 Months of Blogs. In: *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 30--33.

Glance, N., Hurst, M., Tornkiyo, T. (2004). Blogpulse: Automated Trend Discovery for Weblogs. In: *Proceedings of WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*.

Hurst, M. (2006). 24 Hours in the Blogosphere. In: *Proceedings of AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, pp. 73--77.

ICWSM (2007). http://www.icwsm.org/data.html

Kim, J.H. (2005). Blog as an Oppositional Medium? A Semantic Network Analysis on the Iraq War Blogs. In: *Internet Research 6.0: Internet Generations*.

Manning, C.D., Schütze, H. (1999). Foundations of Statistical Natural Language Processing, The MIT Press, London England.

Mei, Q., Liu, C., Su, H., Zhai, C. (2006). A Probabilistic Approach to Spatiotemporal Theme Pattern Mining on Weblogs. In: *Proceedings of the 15th International Conference on World Wide Web, Edinburgh, Scotland*, pp. 533--542.

Oka, M., Abe, H., Kato, K. (2006). Extracting Topics from Weblogs Through Frequency Segments. In: *Proceedings of WWW 2006 Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis, and Dynamics*.

Teng, C.Y., Chen, H.H. (2006). Detection of Bloggers’ Interest: Using Textual, Temporal, and Interactive Features. In: *Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 366--369.

Teng, C.Y., Chen, H.H. (2007). Analyzing Temporal Collocations in Weblogs. In: *Proceeding of International Conference on Weblogs and Social Media*, 303--304.

Thelwall, M. (2006). Blogs During the London Attacks: Top Information Sources and Topics. In: *Proceedings of 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Thompson, G. (2003). Weblogs, Warblogs, the Public Sphere, and Bubbles. *Transformations*, 7(2).