Analyzing Temporal Collocations in Weblogs

Chun-Yuan Teng Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
#1 Roosevelt Rd. Sec. 4, Taipei, Taiwan 106
+886-2-33664888 ext 301
r93019@csie.ntu.edu.tw hhchen@csie.ntu.edu.tw

Abstract
With the popularity of weblogs, it is expected that weblogs contain abundant personal experiences, public opinions, and real events. In this paper, we use the temporal collocations to analyze the term-to-term associations with respect to time in weblogs. We first define a new measure of temporal collocation based on the mutual information, and then conduct experiments using our measure in weblogs. The results reveal that the temporal collocation reflects real-world semantics and real-world events that are happening over time.

1. Introduction
Weblogs are powerful because they allow millions of people to publish and share their ideas easily, and millions more to read and respond. It is desirable to retrieve two kinds of information from weblogs: (1) *the term-to-term association* and (2) *the correlation between the term-to-term association and time*. Among them, *the term-to-term association* can be useful to identify the opinions or positive/negative concerns toward a topic. For example, a weblog article about products may contain different associations of terms such as "expensive price" and "good service" to describe the products. *The correlation between term-to-term association and time* can be helpful for trend analysis and temporal analysis. For example, "President Bush" occurs more frequently than "President Clinton" in the weblogs from 2002 to 2006. Thus, we can observe the famous named entity in specific timestamp.

In this paper, we use *temporal collocation* to model the term-to-term association over time. We modify the pairwise mutual information [1][2] and define the temporal mutual information as follows.

Definition 1 (Temporal Mutual Information) Given a timestamp t and a pair of collocating terms, i.e., x and y, the temporal mutual information can be defined as follows:

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

where $P(x, y \mid t)$ is the conditional probability of co-occurrence of terms x and y in timestamp t, $P(x \mid t)$ and $P(y \mid t)$ denote the conditional probability of occurrence of x and y in timestamps t, respectively.

2. Dataset
In this paper, we use the dataset provided by the ICWSM conference. The weblog data is collected from May 1, 2006 through May 20, 2006. To extract the collocations, we retrieve the collocations within the window of five words. In this way, we get 6,345,173,518 pairs of collocation for our study. To analyze the special events in our dataset, we identify two special events: mother's day (May 14), and the release of Da Vinci Code (May 19).

3. Experiments
In the following, we analyze the temporal collocation from two aspects: representative examples and two special events, i.e., release of Da Vinci Code and Mother's day.

3.1 Representative examples
To analyze the temporal collocation, we provide several examples and observations in this section. Following these examples, we also discuss the behavior of the temporal collocation in different events.

Figure 1 shows two pairs of collocations ("I", "work") and ("go", "church") in the 20 days. We can observe that the temporal mutual information changed periodically in a weekly basis. That is, the temporal mutual information changed repeatedly: starting in the first few days of the week and ending high in the weekend. Besides, this example also shows that the use of pair ("I", "work") is more common than the pair ("go", "church").

Fig. 1: *Two examples of temporal collocation*

3.2 Analysis of events

Figure 2 shows two selected collocations in our dataset, ("with", "mother") and ("mother's", "day"). We can observe that the mutual information of these two pairs reaches the peak near May 14, i.e., Mother's day. The pair ("mother's", "day") is higher than the pair ("with", "mother") containing the stopword "with." The reason is that the occurrences of "with" is very high and the temporal mutual information of ("with", "mother") is significantly reduced under our model.

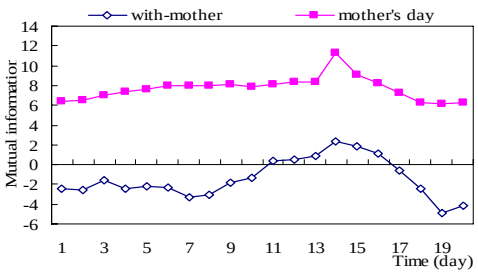

Fig. 2: *Temporal collocation: mother's day*

Figure 3 shows two observations of the other event, i.e., the release of Da Vinci Code. Two examples are provided. "Tom Hank" is the name of the actor; "Da Vinci" is just the fragment of the name of the movie "Da Vinci Code". We can observe some facts from these examples. "Tom Hanks" has higher change of temporal mutual information compared to ("Da","Vinci") near the release date. The reason is that ("Da","Vinci") is discussed more often than ("Tom","Hanks") in the weblogs; however, ("Tom","Hank") is only referred near the release date.

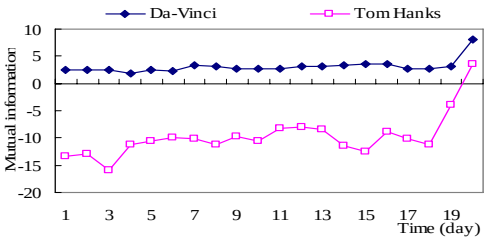

Fig. 3: *Temporal collocation: The release of Da Vinci Code*

4. Conclusions

With the defined temporal mutual information, we analyze the collocation by (1) the collocation in time dimension; and (2) the interesting collocations related to the special events. There are several interesting directions to extend our work. First, we do not consider the multi-word temporal collocations. Second, we also do not consider the factors, such as user, location, and etc. In the future, we may try to model these factors and observe the collocations over time and locations.

Acknowledgements

Research of this paper was partially supported by National Science Council, Taiwan, under the contract NSC95-2752-E-001-001-PAE.

References

[1] Church, K., and Hanks, P. Word Association Norms, Mutual Information and Lexicography. *Computational Linguistics*, 1990, Vol. 16, No. 1, 22-29.

[2] Manning, C. D. and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, London England, 1999.