Identification and Visualization of Emerging Trends from Blogosphere

Makoto Uchida[⋆]
uchida@race.u-tokyo.ac.jp

Naoki Shibata[‡]
shibata@biz-model.t.u-tokyo.ac.jp

Susumu Shirayama[⋄]
sirayama@race.u-tokyo.ac.jp

[⋆⋄]Research into Artifacts, Center for Engineering, the Univ. of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8568 Japan
[‡]School of Engineering, the Univ. of Tokyo, 2-11-6 Yayoi, Bunkyo, Tokyo 113-8656 Japan

Abstract

We propose a method for tracking emerging trends from weblogs, by a hybrid approach of semantic-based and citation-based. In the present method, we form a citation network of the *blogosphere* by regarding entries as nodes and trackbacks as edges, and identify groups (*communities*) of entries by topological clustering on the network. Then we assign featured terms which represent a topic of each community by applying a TF-IDF based technique. The overall results are visualized by graph drawing. Applying the method to a test collection of Japanese weblogs, we confirm that the proposed method works successfully.

Keywords

Topic Identification, Emerging Topic, Emerging Trend, Visualization, Japanese weblog

1. Introduction

We perform an analysis of evolving dynamics of the *blogosphere*, when hot trends are born and saturated. Since contexts of weblogs directly reflect interests and attentions of bloggers, analysis on dynamically-changing contents of weblogs and identification of emerging trends in the *blogosphere* will give us an insight to public interests at a point in time.

There are already proposed some methods to detect bursting of hot terms by a semantic-based approach [1, 3]. In addition to the past researches, we focus on dynamics of community structures, and propose a way to monitor growth of hot trends topologically based on a citation-based approach [2], which identifies communities from a link structure of weblogs. In this paper, we propose a method to identify not only trends which densely-connected entries refer, but also their emergence which change dynamically at each time by a hybrid approach of semantic-based and citation-based.

2. Proposed Method

We construct a network regarding entries as nodes and trackbacks as undirected edges (although trackback is directed, we consider the network undirected, making each edge bidirectional), and we form the network into groups of entries (*communities*). Here, a community is defined as a group of entries within which trackbacks are relatively dense. Note that, community in this definition only focus on a topological aspects of networks, never making use of any contents (titles, body texts, *etc*) in the entries. In this paper, we exploit a hierarchical clustering method proposed by Newman. See Reference [4] for the detail of the method.

Then, in order to identify a topic discussed in each community, we apply TF-IDF technique at two separate steps. First, we calculate scores of terms in a entry by Eqn 1 and regard n largest terms of this score as featured terms of the entry.

$$tfidf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \tag{1}$$

where $tf_{i,j}$ is the number of occurrence of term i in entry j, df_i is the number of entries containing term i, N is the total number of whole entries. This procedure normalizes length of entry bodies and decreases weights of common terms. Second, we calculate scores of terms *in a community* by Eqn (2). This procedure normalizes the sizes of communities.

$$tfidf^*_{i,k} = tf^*_{i,k} \times \log\left(\frac{C}{df^*_i}\right) \tag{2}$$

where $tf^*_{i,k}$ is the number of entries which have the term i in their featured terms in the community k, df^*_i is the number of communities containing term i, C is the number of all communities. The more frequently in a certain community and the less frequently in the other communities a term appears, the larger $tfidf^*_{i,k}$ score becomes. Therefore, terms with larger score can be considered to be featured terms which represent a topic discussed in the community.

3. Testing the method using an empirical dataset

We applied the proposed method to a dataset of Japanese weblogs, collected by tracing trackbacks from randomly chosen, single seed entry. We formed a network of 25,668 entries and 67,828 undirected edges. The data contains URL, published date of the entries, destination and source of trackbacks. They are collected in December, 2004. Applying the Newman's method for community division, the network was divided into 127 communities. By Eqn (1), we extracted 20 terms with largest $tfidf_{i,j}$ score of each entry. Then, by Eqn (2), 20 terms with largest $tfidf^*_{i,k}$ score in each community. The representative results of the top 3 communities in size are shown in the upper rows of Table (1).

In order to verify an advantage of our method, top 5 terms of $tf^*_{i,k}$, without applying Eqn 2, are shown in lower rows of Table 1 for comparison. In these results, scores of general terms and noise terms which appeared in most communities became larger and the scores of proper nouns lower. It can

ID	Size	Top 5 terms of largest $tfidf^*_{i,k}$ (in upper rows), and $tf^*_{i,k}$ (in lower rows)
1	2934	Chuetsu, earthquake (jishin), Niigata pref. (Niigata-ken), victims (hisaisha), Golden Eagles earthquake (jishin), Chuetsu, year (nen), |, 2004
2	2409	Livedoor, baseball team (kyuudan), strike (suto), professional (puro), baseball (yakyuu) |, baseball (yakyu), year (nen), 2004, Livedoor
3	1164	Kimura, Takeshi, annuity (nenkin), weekly (shuukan), public (kouteki) trackback, Kimura, Takeshi, annuity (nenkin), cocolog[9]

Table 1: *Top 5 terms of largest TF-IDF score (upper row) and of most frequent occurrence (of largest $tf^*_{i,k}$) in each community. Originally they are Japanese. Original terms (in Japanese) are shown in brackets together with their translation in English.*

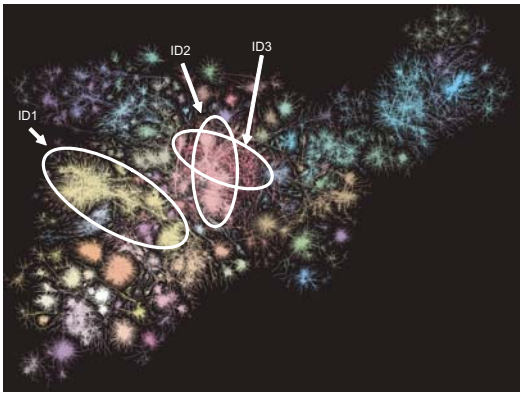

Fig. 1: *(Color online) Visualization and topic mapping to communities of weblog network (See Table 1 together).*

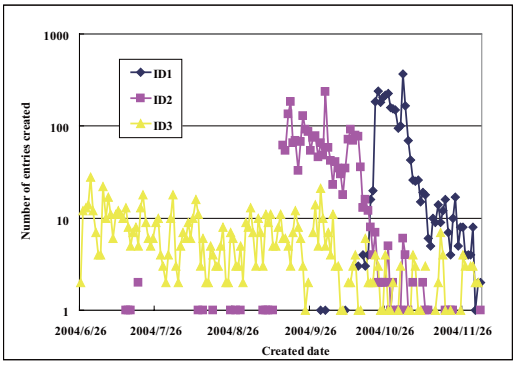

Fig. 2: *Growth pattern of three largest communities (daily number of created entries in a community).*

thus be considered that normalization by Eqn (2) can effectively make unique terms in each community clearer.

The featured terms were confirmed to represent the contents of entries in the communities. In the community ID1, we can see that these entries were about the Mid Niigata Earthquake of 2004, a big earthquake occurred in Niigata Prefecture, Chuetsu district in Japan, by which thousands of people had suffered. ID2 was considered to be a discussion about strike by Japan Professional Baseball Players Association toward the consolidation and new entry of a team. There were many discussions of arguments for and against when a venture business, Livedoor, Ltd., announced to take over an old baseball team. In ID3, bloggers was discussing restructuring public annuity system in Japan. Takeshi Kimura, who was also an influential blogger, had led the discussion.

Those communities are mapped in a visualization of the network as Fig 1. Communities are apparently discrete and varied in sizes, most of which we confirmed to have characteristic topics as well as the three largest communities. Through the visualization, we can see the scale of topics, and connectivities *between* topics could be analyzed.

Growth patterns of communities gave us an insight to emergence of the topics. Time-series number of posted entries in the three largest communities are shown in Fig 2. A significant increase of new entries occurred at about October 20, 2004 on ID1, and at about September 15 on ID2. On the other hand, there seemed to be no peak in ID3. In the real world, The Mid Niigata Prefecture Earthquake of 2004 was on October 23. and strike by Japan Professional Baseball Players Association was on September 17. In the *blogosphere*, upsurge of entries occurred following these real event.

4. Conclusion

We have proposed a method for tracking trend from weblogs, based on social network analysis and linguistic filtering. We formed a network of the *blogosphere* by regarding entries as nodes and trackbacks as edges. Then, we assigned featured terms to each community by a two stage TF-IDF scoring technique to identify a topic about which entries in the community are mainly discussing. Combining a community growth pattern and featured terms, we have showed that the method successfully extracted dynamically-changing emerging trends which correspond to an event in the real world. Moreover, visualization of the communities and the growth of the network has enabled us to grasp intuitively the structure of the *blogosphere*, the communities and their topic, and their scales and growth patterns.

References

[1] K. Balog and M. de Rijke. Decomposing bloggers' moods: Towards a time series analysis of moods in the blogosphere. In *Proceedings of WWE 2006 3rd Annual Workshop on the Weblogging Ecosystem*, 2006.

[2] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspase. In *Proceedings of the 12th international conference on World Wide Web*, pages 568 – 576, 2003.

[3] H. A. Mizuki Oka and K. Kato. Extracting topics from weblogs through frequency segments. In *Pcoreedings of WWE 2006 3rd Annual Workshop on the Weblogging Ecosystem*, 2006.

[4] M. E. J. Newman. Fast algorithm for detecting community structure in netwoks. *Physical Review E*, 69(066133), 2004.