Extracting Topics and Innovators Using Topic Diffusion Process in Weblogs

Tadanobu Furukawa*, Yutaka Matsuo*, Ikki Ohmukai‡, Koki Uchiyama◇, Mitsuru Ishizuka*
furukawa@mi.ci.i.u-tokyo.ac.jp, matsuo@bizmodel.t.u-tokyo.ac.jp, i2k@nii.ac.jp, uchi@hottolink.co.jp, ishizuka@i.u-tokyo.ac.jp

*University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 135-8656, Japan
‡National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
◇Hotto Link Inc., 1-6-1 Otemachi, Chiyoda-ku, Tokyo 100-0004, Japan

Abstract

The diffusion process on weblogs has attracted great interest since the early days of weblog studies. We propose a ranking technique which extracts topics and innovators by analyzing that process. Our method identifies URLs of topics and the bloggers who trigger topic diffusion. Our assumption is that the strength of propagation of a topic is determined by the influences of topics and bloggers. This decomposition is attained through singular value decomposition (SVD); We construct a *diffusion matrix*: the first left and right singular vectors are regarded respectively as the influences of the topics and the bloggers. We show that our method can extract propagative topics (which is not bursty because of the media effect) as well as the influential bloggers.

Introduction

Weblogs (blogs), as a social medium, have been analyzed to understand the process of information diffusion on the Web. When bloggers find interesting topics in other weblogs, they write a post referring to the article, thereby causing further propagation on the blogosphere. To date, several studies have provided overviews and insights related to the diffusion process on weblogs (Adar & Adamic 2005; Zhao, Mitra, & Chen 2007). Trend detection has been attempted to understand the trends of the blogosphere (Chi, Tseng, & Tatemura 2006). This study combines the approaches used in the two streams of research and proposes a topic and innovator extraction method based on the diffusion process of weblogs.

Our approach differs from that of other topic extraction. We extract URLs as topics based on the diffusion process on weblogs. For example, if URLs such as *Google* and *Yahoo* are prominent in weblogs, then we do not want to extract these URLs. They might simply be more frequent, but might not accurately reflect propagation in weblogs.

In our algorithm, we first use information from a blogger's reading behavior. We then construct a matrix based on whether the blogger reads other blogs before using the URL. Then, we apply singular value decomposition (SVD) to the matrix. The first left and right singular vectors are regarded as the respective influences of the URLs and the bloggers. We show that our method can identify propagative topics and influential bloggers. We use the database of a blog-hosting service in Japan called *Doblog*[1]. This dataset consists of 1,540,077 entries by 52,525 users from October 2003 to June 2005.

The remainder of the paper is organized as follows. We explain the definition of diffusion and explain our method for topic extraction in Section 2. Section 3 presents an explanation of the experimental results. Finally, we describe future work and conclude this paper.

Topic and innovator extraction from diffusion

For this study, we define a URL as a topic and extract influential topics and bloggers from propagation of URLs. We consider the propagation is determined both by *influence of the topic* and *influences of bloggers*. The diffusion is defined by bloggers' reading and writing posts because our previous work shows that regularly reading relations with other blogs tend to convey information that is interesting to particular bloggers with higher probability (Furukawa *et al.* 2007).

We define the diffusion of a URL among bloggers as follows: *a blogger read other bloggers' posts, including the URL and then writes a post including the URL.* The diffusion of a URL R from a blogger B_i to a blogger B_j is therefore defined as: (1) B_j regularly reads [2] B_i's blog. (2) Then, B_i creates a post including a URL R. (3) Subsequently, (2) B_j creates a post including a URL R. Thus by this definition, we concentrate solely on the influence of Doblog's blogs, rather than other information sources such as news sites and external blogs, on the URL diffusion process.

We represent the topic diffusion as a *diffusion matrix* A (Eq. 1). This matrix represents the extent to which each URL is propagated by each blogger. In matrix A, for example, a_{12} denotes the degree to which the blogger $blogger_2$ diffused the topic URL_1. The degree is calculated as follows: d_s denotes the days used for the convection of URL_i from $blogger_j$ to $blogger_s$; also, S denotes the adopters of diffusion from a $blogger_j$. Consequently, $a_{ij} = \Sigma_{s \in S} (\log d_s)^{-1}$. This representation incorporates

[1] Doblog (`http://www.doblog.com/`), provided by NTT Data Corp. and Hotto Link, Inc.

[2] Visiting more frequently than every 10 times that the blogger logs in to the doblog system.

the characteristics of the influence of topics/bloggers in column/row vectors.

$$A = \begin{array}{c} \\ URL_1 \\ \vdots \\ URL_m \end{array} \begin{array}{c} \begin{array}{ccc} blogger_1 & \cdots & blogger_n \end{array} \\ \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \end{array} \quad (1)$$

We assume that unique influences of topics exist ($\vec{P} = (p_1, \cdots, p_m)$) and bloggers ($\vec{Q} = (q_1, \cdots, q_n)$), where n and m represent the size of topics and blogs; we also assume that p_i and q_j represent the influences of a topic and a blogger, and that they determine the diffusion matrix A. Consequently, we extract these two vectors and maximize the approximation of the matrix.

The SVD for an approximation of matrix X is given as $\tilde{X} = U\tilde{\Sigma}V^T$ (U, Σ, V are matrices). When $\mathrm{rank}(\tilde{X}) = 1$, the matrices U and V are the vectors and $\tilde{\Sigma}$ is a scalar. The product $U\tilde{\Sigma}V^T$ is the best approximation of X by a vector product. The approximation by SVD (rank $= 1$) for the diffusion matrix A returns the row vector $U_{m\times 1}$ as $\vec{P}$ and column vector $V_{1\times n}$ as $\vec{Q}$. We assume the row-wise and column-wise characteristics of diffusion matrix as the influences of topics and bloggers. For that reason, these singular vectors are used as the influences in this paper. The URL of a high value in $\vec{P}$ is identified as an influential topic. The blogger of high value in $\vec{Q}$ is extracted as an innovator who adopts a new topic earlier than others.

Experiment

We evaluate the proposed method according to two tasks: ranking of topics and innovators. We applied our method to a doblog dataset. We use data of 2,648 randomly chosen bloggers, including the 150 URLs that are cited more than five times in their posts.

Topic ranking

The first task is to order the influential URLs (= topics) higher. We separate the URLs into two classes: a *positive* class including the 30 web sites that seems to be diffused through reading channels, such as individual blog posts; and a *negative* class including the 120 web sites, such as company or movie sites, which seem to be diffused by other information sources.

Table 1 presents the top seven URLs that were ordered using our method. The top places are almost entirely occupied by the sites that attract interest of readers such as individual blog posts taking a vote or fortune-telling sites. In contrast, the URLs of a product or a movie, such as *iPod* and *Spiderman*, are cited frequently, but they do not cause diffusion well and the ranks are low. We calculate their *pairwise accuracy* to evaluate the quality of this ranking. If $O(x)$ is the ranking of URL x, and $H(x)$ is the correct ranking (positive/negative) assigned manually, then $pairwise\ accuracy = \frac{|H_p \cap O_p|}{|H_p|}$ ($O_p = \{x, y : O(x) > O(y)\}, H_p = \{x, y : H(x) > H(y)\}$). The accuracy is apparently around 76%, which means that the URLs in a positive class are likely to be highly influential.

Table 1: Ordering of URLs. A URL with * is assigned as the negative class.

rank	contents of URL
1	Introduces blogger's favorte movies.
2	Introduces blogger's favorte music albums.
3	Predicts one's personality.
4	* Provides technology related news. (*Slashdot*)
5	Checks the traffic load on Doblog.
–	pairwise accuracy = 76.5%

Table 2: Comparison of ordering of bloggers to the ranking of number of trackbacks/comments per day.

blogger ID	proposed	trackback	comment
A	1	12	406
B	2	30	174
C	3	39	207
D	4	8	33
E	5	283	177

Innovator ranking

A feature of our method is that the influence of bloggers can also be extracted concurrently with extraction of the influence of topics. The ranking of bloggers is presented in Table 2. The influential bloggers are expected to be ranked in a high position. Therefore, we compare the ranking of proposed method with those of the daily number of comments/trackbacks. Considering the overall number of blogs to be 52,525, the top ranks of the proposed method collect quite numerous comments and trackbacks. They are the innovators who write posts about new topics earlier and interact with other bloggers more frequently than usual. Our algorithm identifies them as highly influential bloggers.

Conclusion

This study has examined a method to extract influential topics and innovators from a topic diffusion process. We assumed that the topic diffusion consists of influences of topics and bloggers. This definition facilitated ordering of topics that attract bloggers through reading a post in a high rank, and, simultaneously facilitated identification of highly interactive bloggers.

For future work, we plan to elucidate characteristics of diffusion in the blogosphere. Proper modeling of the topic diffusion process, considering features of topics and bloggers, will improve the accuracy of our algorithm.

References

Adar, E., and Adamic, L. A. 2005. Tracking information epidemics in blogspace. In *Web Intelligence 2005*.

Chi, Y.; Tseng, B. L.; and Tatemura, J. 2006. Eigen-trend: Trend analysis in the blogosphere based on singular value decompositions. In *Proc. CIKM 2006*.

Furukawa, T.; Matsuo, Y.; Ohmukai, I.; Uchiyama, K.; and Ishizuka, M. 2007. Social networks and reading behavior in the blogosphere. In *ICWSM 2007*.

Zhao, Q.; Mitra, P.; and Chen, B. 2007. Temporal and information flow based event detection from social text streams. In *Proc. AAAI-07*.