Construction of a Blog Network Based on Information Diffusion

Seung-Hwan Lim
Dept. of Electronics and
Computer Engineering
Hanyang University
South Korea
shlim@agape.hanyang.ac.kr

Sang-Wook Kim
Dept. of Electronics and
Computer Engineering
Hanyang University
South Korea
wook@hanyang.ac.kr

Soyoun Kim
School of Information and
Communications
Hanyang University
South Korea
kimsoyoun@hanyang.ac.kr

Sanghyun Park
Computer Science
Department
Yonsei University
South Korea
sanghyun@cs.yonsei.ac.kr

ABSTRACT

The blog world is a representative online society. To understand the nature of the blog world, there have been many research efforts on analyzing information diffusion and blogger activities. The *independent cascade model* is appropriate to explain information diffusion in the blog world. For the model to be employed, the blog world should be represented as a form of a network. For accurate analysis, it is crucial to assign a *diffusion probability* to each edge between a pair of bloggers in the blog network. In this paper, we propose a novel method to assign a diffusion probability to an edge for a pair of bloggers that reflects well the phenomenon of actual information diffusion between them. We verify the superiority of our approach by performing extensive experiments with real-world blog data.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: SOCIAL AND BEHAVIORAL SCIENCES—*Sociology*; H.2.8 [**Database Management**]: Database Applications—*Data mining*

General Terms

Algorithms, Human Factors, Experimentation

Keywords

Data Mining, Social Network Analysis, Blogs, Information Diffusion

1. INTRODUCTION

A *social network* is a kind of a network that consists of members and relationships in a society. Analyzing various characteristics of a social network is called a *social network analysis (SNA)* [22]. For a long time, most of social network data contained only the information regarding the *presence* of relationships between pairs of members. This causes researches on social network analysis to focus on the topological characteristics of social networks [8][10][16][19][20][23].

With the development of the Internet and Web technology, however, social networks emerged online. Online social networks provide information much richer than traditional ones because every change in them is easily maintained in a database. In online social networks, the *strength* of a relationship, which could be different with different pairs of members, can be easily captured. So, recent studies are focusing on deriving more detailed and accurate results by analyzing the relationship information [1][12].

A blog service is one of the most widely-used services based on online social networks. A *blog* is a type of a personal website in which bloggers can write their own thoughts and opinions in a *post* [3][4][7][17][18][21]. Bloggers influence one another through their postings and perform various activities in the blog world. A variety of relationships are also established between a pair of bloggers there. The network consisting of bloggers and their relationships is called *blog network*.

Blog service companies provide convenient functions to bloggers such as copying someone else's post to their blogs or writing a new post related to someone else's post in their own blogs in case they are interested in the post. These functions are called *scrap* and *trackback*, respectively, which cause *information diffusion* over the blog world. It is an interesting research issue to analyze such information diffusion for understanding how to detect abused information diffusion, how to build successful marketing strategies for bloggers, and how to revitalize blog networks [5][6][13][14][15][24].

In order to analyze information diffusion phenomena in the blog world, the *independent cascade model* [6] can be accepted. To apply this model, the blog world needs to be represented as a blog network. A blog network, which is built based on information diffusion history, is composed of

SAC'11 March 21-25, 2011, TaiChung, Taiwan.
Copyright 2011 ACM 978-1-4503-0113-8/11/03 ...$10.00.

nodes and edges, each of which expresses a blogger and a relationship between a pair of bloggers, respectively. For analysis, the independent cascade model requires a diffusion probability for every edge. Information diffusion in the blog world is captured correctly only when the diffusion probability is assigned reflecting well the strength of a relationship between the corresponding pair of bloggers. However, most of previous works assumed these probabilities are given in advance, and did not deal with how to obtain the probabilities. Many of them simply assigned an identical value to all the edges within a network for analysis.

In this paper, a novel method to construct a blog network is proposed for analyzing information diffusion in the blog world. Once the network is constructed, the independent cascade model can be applied to understand how the information is diffused over the blog world. An algorithm to compute a diffusion probability to each edge between a pair of bloggers is proposed by carefully analyzing the history of information diffusion. We demonstrate the effectiveness of the proposed method by performing extensive experiments using real-world blog data.

The paper is organized as follows. Section 2 introduces the characteristics of the blog world. Section 3 briefly reviews previous studies on information diffusion in social networks. Section 4 proposes our method and discusses its characteristics. Section 5 presents and analyzes experimental results. Section 6 summarizes and concludes the paper.

2. BLOG NETWORK

The environment of blog services can be summarized as follows. Blog service companies provide bloggers with a function called *bookmarks* or *neighbors*, which helps to add some blogs to her/his favorites-list that makes it possible for her/him to visit those blogs easily with a single click [3][7][17][18][21]. A blogger can also perform actions such as *read*, *comment*, *trackback*, and *scrap* on a post in someone else's blog [3][18][21].

Trackback and scrap can be viewed as a way of *reproducing* the original post. The reproduced post may in turn trigger someone else to perform actions such as read, comment, trackback, and scrap. As a result, these two actions cause information created by bloggers to be diffused over the blog world.

Relationships among bloggers can be expressed as edges while bloggers are represented by nodes. So, the blog world is characterized as a form of a network. When blogger B adds another blogger A to one's own bookmark or takes actions such as read, comment, trackback, and scrap on the post of blogger A, these actions can be considered as a consequence of blogger B being influenced by blogger A. Thus, we can establish edges for constructing a blog network in the following two different ways.

First, an edge can be formed between bloggers who are connected by *bookmarks.* This is based on the assumption that a bookmark is formed only when a blogger is influenced by another blogger. Due to the static nature of a bookmark, however, this may not reflect *current* influential relationships between bloggers. Second, an edge can be formed between a blogger who has reproduced a post and its original owner of the post. This is based on the assumption that a blogger does reproducing actions such as trackback and scrap on a post when she/he is influenced greatly by the blogger who is the owner of the post. This way of edge formulation differs from the first one because it reflects *recent* influence between bloggers when we consider actions performed within the recent time window.

Figure 1 depicts an example of a blog network. Figure 1(a) shows an example of the blog world. Figure 1(b) shows a blog network formed by *bookmarks* in Figure 1(a). Figure 1(c) shows another blog network formed by *reproductive actions* in Figure 1(a).

In Figure 1(a), large rectangles labeled $A \sim E$ represent bloggers while small round rectangles within a large rectangle represent posts created by a blogger. The dotted arrows indicate bookmarks among bloggers while the lined arrows indicate actions performed by bloggers. The posts with the same color represent those related through reproductive actions.

In Figures 1(b) and 1(c), a circle represents a blog, and an arrow does the influence relationship between bloggers. In Figure 1(b), $B \rightarrow A$ indicates that blogger A added blogger B in her/his bookmark, which implies that blogger A was influenced by blogger B. In Figure 1(c), $C \rightarrow A$ indicates that blogger C performed a trackback to post 3 of blogger A, which exhibits the influence of blogger A over blogger C.

3. RELATED WORK

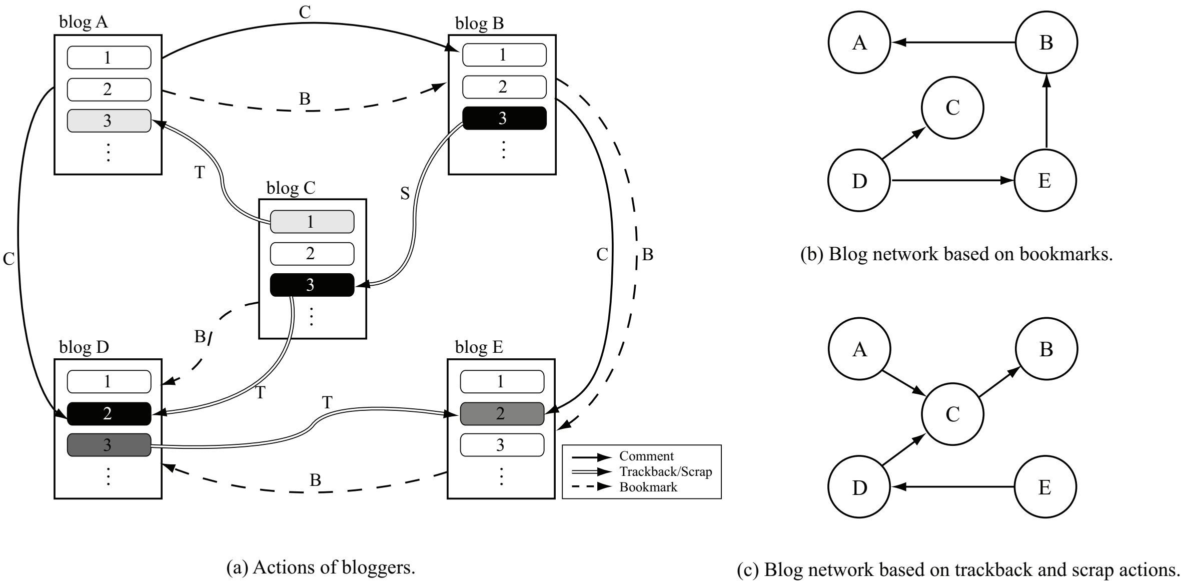
Models for information diffusion in social networks include the linear threshold model [9], the independent cascade model [6], and the generalized cascade model [11]. The common idea of these models is the following. Nodes may influence one another and therefore a node which was influenced by another node may have characteristics similar to those of the influencing node. In this case, we say the influenced node is *assimilated* by the influencing node.

In [9], the *linear threshold model* was proposed. The linear threshold model designates a threshold value to each node and a weight to the relationship between nodes. When a specific node's accumulated influence received from surrounding nodes is greater than its threshold value, it is regarded as assimilated by those surrounding nodes. However, the linear threshold model is inappropriate to be applied to a blog network. It is because, while information diffusion occurs through the *independent* relationships among bloggers in the blog world, the linear threshold model calculates the total influence to a node by *adding up* the weighted influences from its neighboring nodes.

[6] proposed the *independent cascade model.* The independent cascade model designates a probability to the relationship between nodes, and assimilation decision is made based upon this. We will call this probability *assimilation probability between nodes.* The reasoning behind the model is that a blogger diffusing a certain post in the blog world is not because of the influence from her/his neighbors but because of the influence from a *single* blogger who possesses the post. Therefore, the independent cascade model is appropriate to analyze information diffusion in the blog world.

In [11], a *general cascade model* is proposed. The general cascade model eliminates the condition in the independent cascade model that, in order to assimilate a specific user, the neighboring users influence him independently, thus generalizing the characteristics of the linear threshold model and the independent cascade model.

Due to the reasons mentioned above, we decided to employ the independent cascade model to explain information diffusion in the blog world. For analysis by using the independent



(a) Actions of bloggers.

(b) Blog network based on bookmarks.

(c) Blog network based on trackback and scrap actions.

Figure 1: An example of a blog network.

cascade model, transforming a given blog world into its corresponding blog network needs to be preceded. In addition, every edge between a pair of bloggers requires a diffusion probability to construct a blog network. To correctly understand information diffusion in the blog world, this probability should be accurately computed reflecting actual information diffusion phenomenon well. Previous studies, however, mostly focused on proposing the models that explain information diffusion phenomenon in social networks. They assumed that the diffusion probabilities are given from applications or simply assigned an identical value to all the edges for analysis. In this case, the information diffusion cannot be analyzed correctly because it is determined only by the topology of the network.

4. PROPOSED METHOD

4.1 Terms and symbols

Table 1 lists the terminologies and symbols used for further discussions. U_A represents blogger A. D_A represents a collection of posts that U_A creates and possesses while $D_{A,i}$ represents post i of U_A. $D_{A\to B}$ is a collection of posts that are diffused by U_B from D_A. $P_{A\to B}$ is the probability that a post in D_A is diffused by U_B. $score(D_{A,i})$ indicates the score of blogger U_A's intention of diffusing post $D_{A,i}$, which is used to calculate diffusion probabilities in the following section. To compute $score(D_{A,i})$, other bloggers' actions such as write, read, comment, trackback, and scrap, denoted as W, R, C, T, and S, respectively, are used. In order to assign different importance to different types of actions, weights, denoted as W_W, W_R, W_C, W_T and W_S, are assigned.

4.2 Basic idea

Our method is motivated by the observation that $P_{A\to B}$, the probability of a post of U_A being diffused to U_B, is proportional to the number of posts that are diffused by U_B from D_A and also inversely proportional to the number of posts in D_A created by U_A. So, probability $P_{A\to B}$ is basically computed as in Equation (1).

$$P_{A\to B} = \frac{|D_{A\to B}|}{|D_A|} \tag{1}$$

In Equation (1), $|D_A|$ and $|D_{A\to B}|$ indicate the numbers of posts in D_A and $D_{A\to B}$, respectively. $P_{A\to B}$ refers to the ratio of the number of posts in $D_{A\to B}$ to the number of posts in D_A.

4.3 Improvement

Normally, bloggers create posts with two different intentions. The first one is to provide useful information for other bloggers. In this case, the owner expects that her/his post will be diffused over the blog world and thus will influence other bloggers. The second one is to keep her/his private thoughts and emotions for archival purposes. In this case, she/he does not want the post open to other bloggers. So, it would not be reasonable if we include those posts belonging to the second category in $|D_A|$ of Equation (1).

To solve this problem, we should understand the owner's intention when she/he is writing a post. Direct inquiries to owners on the intention for their posts could be accurate but infeasible in real situations. Thus, we propose to estimate the author's intention on a post by quantifying other bloggers' actions on the post. This is based on the fact that a post created by the intention of diffusion tends to incur more actions of other bloggers than a post created by the other intention.

We define the score of a blogger's intention of a post being diffused as a degree of how much the blogger intends to diffuse the post in question. This is derived from the amount of actions induced by other bloggers on the post. Equation (2) shows a formula to compute this score. The $score(D_{A,i})$ indicates the score of U_A's intention of post i being diffused when she/he created the post. This score is computed by

Table 1: Terminologies and symbols

Symbols	Definitions
U_A	Blogger A
$D_A = \{D_{A,1}, D_{A,2}, \cdots\}$	A collection of posts that U_A possesses
$D_{A,i}$	Post i of U_A
$D_{A\to B}$	A collcction of posts that were diffused by U_B from D_A
$P_{A\to B}$	Probability that a post in D_A is diffused by U_B
$score(D_{A,i})$	Score of blogger U_A's intention of diffusing post $D_{A,i}$
W, R, C, T, S	Types of bloggers' actions
W_W, W_R, W_C, W_T, W_S	Weights of bloggers' actions

the weighted sum of all the actions, i.e., read, comment, trackback, and scrap.

$$\begin{aligned} score(D_{A,i}) &= W_R \times R\_Count(D_{A,i}) \\ &+ W_C \times C\_Count(D_{A,i}) \\ &+ W_T \times T\_Count(D_{A,i}) \\ &+ W_S \times S\_Count(D_{A,i}) \end{aligned} \tag{2}$$

We denote D_A^* as a subset of D_A containing only those posts created by the intention of diffusion. We consider a post in D_A to belong to D_A^* if its score is greater than a given threshold θ. As a result, Equation (1) is refined as Equation (3).

$$P_{A\to B} = \frac{|D_{A\to B}|}{|D_A^*|}$$

$$\text{where } D_{A,i} \in D_A^* \text{ if } score(D_{A,i}) > \theta \tag{3}$$

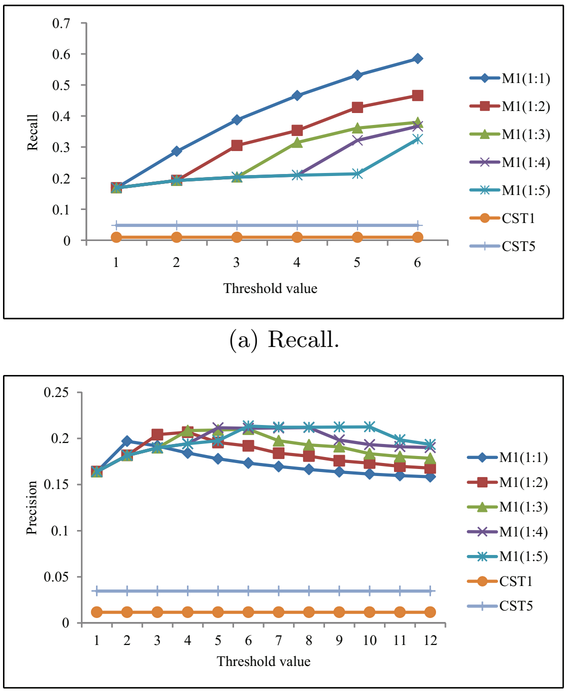
5. PERFORMANCE ANALYSIS

For experimental analysis, we used anonymized data collected from blog.naver.com, one of the largest blog services in Korea. In building a blog network, we established an edge between two bloggers by using the information diffusion history between them rather than using their bookmarks.

For performance comparisons, we ran the following three different methods for analysis: (1) M1(in Section 4.3): our method of computing a diffusion probability with the number of posts created by diffusion intention, (2) CST1: a method of assigning constantly 1% to every edge, and (3) CST5: a method of assigning constantly 5% to every edge. The last two are those employed in [11].

To evaluate the performance of a method, we compared the following two information diffusion histories: the actual diffusion information history found in the blog world and the generated diffusion history obtained by applying the independent cascade model to a blog network built by the method. We employed *recall* and *precision*, widely used for measuring accuracy in the information retrieval field [2]. We obtained recall (and precision) averaged over all the actual diffusion histories in our blog data.

In the experiment, we analyzed the performance by changing the ratio of the action weights of comment and trackback/scrap as 1:1, 1:2, 1:3, 1:4, and 1:5, respectively. We set trackback and scrap to have the same weight because these two actions have similar characteristics of reproduction. We started the threshold of M1, which is the criteria for determining a post to be created by the intension of diffusion, from 1 and increased it in step of 1. Figure 2 shows the results of the experiment. The x-axis represents the threshold value, and the y-axis represents recall and precision in Figures 2(a) and 2(b), respectively.



(a) Recall.

(b) Precision.

Figure 2: Performance of M1 with different action weights and thresholds.

Figure 2(a) demonstrates that recall increases as the action weight ratio of comment and trackback/scrap decreases and also as the threshold increases. This can be explained as follows. As the action weight ratio decreases and the threshold value increases, the number of posts considered as being created by the diffusion intention decreases. This makes a high diffusion probability to be assigned to edges, and consequently causes information to be diffused more widely over a blog network. As a result, more bloggers are included in the information diffusion history made by M1, which increases the possibility of including those bloggers in the actual information diffusion history. With the action weight ratio of 1:1, we observe that M1 improves the recall of CST1 and CST5 by 18∼62 times and 4∼12 times, respectively.

In Figure 2(b) we can see that the precision of M1 also increases with the increase of the action weight ratio. Also, as threshold increases, the precision of M1 increases to a point, but decreases after the point. With the higher action weight ratio, the inclination of the increase and decrease becomes small, which, in turn, implies the threshold influences precision less with a larger action weight ratio. A lesson from the experiment is that, when an analyst sets the action weight ratio and the threshold to analyze information diffusion accurately, she/he should adjust the action weight ratio and the threshold simultaneously rather than one by one. In case the action weight ratio is set to 1:1, we see that M1 shows precision higher than CST1 and CST5 by 14∼17 times and 5∼6 times, respectively.

6. CONCLUSIONS

In this paper, we have discussed a method of constructing a blog network in order to analyze information diffusion in the blog world. For successful analysis of the information diffusion using this model, accurate assignment of diffusion probability to each edge in the blog network is crucial. In this paper, we have proposed a method of computing a diffusion probability for each edge. Our method is based on the observation that the diffusion probability between two bloggers is proportional to the number of posts diffused by the influenced blogger, and is also inversely proportional to the number of posts created by the influencing blogger. So, the proposed method basically computes the probability based on their ratio. For more accuracy, we have used a notion of a blogger's intention for a post being diffused and have proposed a way of its quantification. We have shown the superiority of the proposed method by performing experiments with various experimental settings.

7. ACKNOWLEDGMENTS

This work was supported by NHN Corp. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor. This work was also supported by the Mid-Career Researcher Program through the NRF (National Research Foundation) grant funded by the MEST (Ministry of Education, Science, and Technology) (Grant No. 2008-0061006) and the MKE (Ministry of Knowledge Economy) under the National HRD Support Program for Convergence Information Technology supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2010-C6150-1001-0005). We would like to thank Mr. Kyu-Hwang Kang and Mr. Young-Joo Do for their efforts on implementing our method and also performing extensive experiments.

8. REFERENCES

[1] L. Adamic, O. Buyukkokten, and E. Adar. A Social Network Caught in the Web. *First Monday*, 8(6):1–22, 2003.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, 1999.

[3] Daum Communications Corp. Daum blog. `http://blog.daum.net/`.

[4] Daum Communications Corp. Tistory. `http://www.tistory.com/`.

[5] G. Ellison. Learning, Local Interaction, and Coordination. *Econometrica: Journal of the Econometric Society*, 61(5):1047–1071, 1993.

[6] J. Goldenberg, B. Libai, and E. Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, 12(3):211–223, 2001.

[7] Google. Blogger. `http://www.blogger.com/`.

[8] M. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):7821–7826, 1973.

[9] M. Granovetter. Threshold Models of Collective Behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[10] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabási. The Large-Scale Organization of Metabolic Networks. *Nature*, 407(6804):651–654, 2000.

[11] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the Spread of Influence Through a Social Network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[12] R. Kumar, J. Novak, and A. Tomkins. Structure and Evolution of Online Social Networks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 611–617, 2006.

[13] Y. Kwon, S. Kim, and S. Park. An Analysis of Information Diffusion in the Blog World. In *Proceeding of the 1st ACM International Workshop on Complex Networks Meet Information & Knowledge Management*, pages 27–30, 2009.

[14] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical Properties of Community Structure in Large Social and Information Networks. In *Proceeding of the 17th International Conference on World Wide Web*, pages 695–704, 2008.

[15] X. Li, Y. Wang, and A. Acero. Learning Query Intent from Regularized Click Graphs. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 344–345, 2008.

[16] S. Milgram. The Small World Problem. *Psychology Today*, 2(1):60–67, 1967.

[17] MySpace Inc. Myspace.com. `http://www.myspace.com/`.

[18] NHN Corp. Naver blog. `http://blog.naver.com/`.

[19] M. Nowak and R. May. *Virus Dynamics: Mathematical Principles of Immunology and Virology.* Oxford University Press, 2000.

[20] S. Redner. How Popular Is Your Paper? An Empirical Study of the Citation Distribution. *The European Physical Journal B*, 4(2):131–134, 1998.

[21] SK Communications Corp. Cyworld. `http://www.cyworld.com/`.

[22] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications.* Cambridge University Press, 1994.

[23] D. Watts. *Small Worlds: The Dynamics of Networks Between Order and Randomness.* Princeton University Press, 2003.

[24] WegoNet. Brand Strategy in Communities. *E-Design*

Press, 2004.