

The Information Diffusion Model in the Blog World

Yong-Suk Kwon
Dept. of Electronics and
Computer Engineering
Hanyang University
Seoul 133-791, Korea
neo@zion.hanyang.ac.kr

Sang-Wook Kim
Dept. of Electronics and
Computer Engineering
Hanyang University
Seoul 133-791, Korea
wook@hanyang.ac.kr

Sunju Park
School of Business
Yonsei University
Seoul 120-749, Korea
boxenju@yonsei.ac.kr

Seung-Hwan Lim
Dept. of Electronics and
Computer Engineering
Hanyang University
Seoul 133-791, Korea
shlim@zion.hanyang.ac.kr

Jae Bum Lee
NHN Corp.
Seongnam 266-1, Korea
jblee@nhn.com

ABSTRACT

In the blog network, the posts in a blog can be diffused to other blogs through trackbacks and scraps. Analyzing information diffusion in the blog network is an important research issue that can be used for predicting information diffusion, detecting abnormality, marketing, and revitalizing the blog world. Existing studies on information diffusion in a blog network define explicit relationships between blogs and analyze the word-of-mouth effect through such explicit relationships only. However, it has been observed that more than 85% of all information diffusion in a blog network occurs through non-explicit relationships. In this paper, we propose a new model that considers both the explicit and non-explicit relationships between blogs in order to explain the information diffusion phenomena in a blog network. We add a super node and the relationships between the super node and blogs as broadcast edges and register edges to the existing information diffusion model and assign the assimilation probability to every relationship. The expanded information diffusion model improves the accuracy of the basic model by taking into account the degrees of diffusion powers of posts. We verify the superiority of the proposed model through extensive experiments of information diffusion at a real blog network. The experimental results show that our expanded information diffusion model generates 77% less errors than the existing model.

Categories and Subject Descriptors

J.4 [Computer Applications]: SOCIAL AND BEHAVIORAL SCIENCES—*Sociology*; H.3.3 [Information Systems]: INFORMATION STORAGE AND RETRIEVAL—*Retrieval models*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

The 3rd SNA-KDD Workshop '09 (SNA-KDD'09), June 28, 2009, Paris, France.

Copyright 2009 ACM 978-1-59593-848-0 ...\$5.00.

General Terms

Human Factors, Algorithm, Experimentation, Economic

Keywords

Social Network Analysis, Blog, Data Mining, Information Diffusion, Information Diffusion Model

1. INTRODUCTION

A *social network* is a way of expressing a group of people within some society and the relationships among those group members as a network. Most existing studies on social network have carried out analysis on topological characteristics of social networks [6, 17, 25, 28, 29, 31, 32]. Analyzing the social network and deriving various characteristics of that society is called *social network analysis* (SNA) [30]. With the development of the Internet, social networks have appeared online. Online social network data is different from existing social network data in that they include richer information, such as the degrees of relationships between group members. Accordingly, compared to previous studies that use information on whether or not relationship exist between members, studies on online social networks are developing and exploring many different approaches to utilize these additional information [2, 3, 23].

Blog worlds are a representative example of online social networks. A *blog* is a personal website on which a *logger*, the owner of the blog, can write his or her thoughts online [10, 14, 19, 27]. In a blog network, a blogger is provided with a function called *blogroll* that enables him or her to maintain a relationship with other bloggers either to obtain needed information or to maintain a friendly relationship. This function is similar to the shortcut function of web browsers (such as favorites list or hyperlinks), and it performs the function of helping one to easily visit the blogs that one wants anytime. In this paper, a relationship formed between blogs through function *blogroll* is defined as an *explicit relationship*.

The *trackback* is a function that allows bloggers to make their posts linked to other blogger's posts [7, 26]. The *scrap* function, supported by most Korea's blog service companies [10, 14, 19, 27] is a function that allows bloggers to copy

another bloggers' posts to their blogs. We define *information diffusion* as a phenomenon that a post is disseminated in a blog world through trackback or scrap. Analyzing such information diffusion is a useful research issue, which can be used for predicting information diffusion, detecting abnormality, marketing, and revitalizing the blog world [13, 16].

Existing studies on information diffusion in a blog world have mainly focused on explaining information diffusion through the *word-of-mouth effect* of between the social network members with explicit relationship [8]. As bloggers visit the blogs in their blogrolls and diffuse post (by trackbacking or scraping the post of their interest), such approach can be seen as appropriate. However, our analysis has revealed that information diffusion within a blog world does not necessarily occur through explicit relationships. For example, a post that is introduced on the main page of the blog service company can be explosively diffused to the blogs with no prior relationship.

Therefore, this paper proposes an information diffusion model that can explain both the information diffusion arising from the word-of-mouth effect between bloggers with explicit relationship and those which cannot be explained with it. Using the existing *independent cascade model* [16] as a basis, we propose a new model with additional elements. The new elements here refer to the *virtual space* (i.e., the main page of the blog service provider) that exposes posts to a lot of bloggers with no prior relationship, and the *relationship* between this space and each of the bloggers. In addition, we propose an extended model to increase the accuracy of the new model. Finally, we verify the accuracy of the proposed approach through experiments with real-world blog data.

This paper is structured as follows. Section 2 introduces a blog world in detail and discusses the importance of information diffusion in the blog world. Section 3 describes related work. Section 4 explains the information diffusion model being proposed in this paper. In Section 5, the proposed technique and the existing techniques are used to analyze an actual blog network, and their results are compared. Section 6 concludes the paper and suggests a direction for future research.

2. MOTIVATION

The blog world is an online social network composed of blogs [10, 14, 19, 27]. A *blog* is a type of a personal website where personal thoughts or opinions can be recorded as online documents. Online documents that record these personal thoughts and opinions on blogs are called *posts*.

Similar to the *hyperlink* of the web, *blogroll* establishes an explicit relationship with another blog that the blogger is interested in. Through these established relationships, a blogger easily moves to any of the relevant blogs at any time.

Bloggers' activities regarding posts can be largely divided into two types. One type of activities is what can be done on one's own blog, such as reading one's own posts or composing posts. The other type of activities is what can be done on someone else's blogs, such as reading the posts of others, leaving one's own opinions as comments, *scraping* (copying) the posts to one's own blog, or trackbacking by putting a link to other posts, which leads to one's own blog where one's opinions about the same topic have been written and recorded.

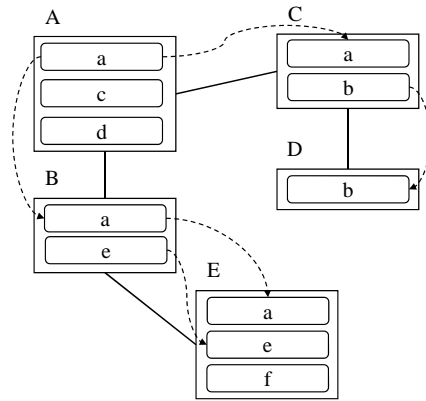


Figure 1: Blog world example.

When a blogger trackbacks or scraps a post, it can be seen as the result of the blogger judging that that post is useful to his or her blog. Moreover, because the post has been trackbacked or copied to one's blog, it also becomes available for other bloggers who visit the blog. Accordingly, the act of trackbacking or scraping a post is an important and distinctive action that explains *information diffusion* in the blog world.

Figure 1 shows an example of the blog world. The rectangles *A-E* represent blogs, and the small rectangles within blogs *a-f* represent posts. The solid lines between blogs show explicit (blogroll) relationships between blogs. The arrows between posts show the posts which have been trackbacked or scrapped to other blogs. In case of post *a*, it was composed on blog *A*, and was scrapped by bloggers *B* and *C* and copied to their blogs.

Information in the blog world is recorded as posts. Accordingly, information diffusion in the blog world occurs through dissemination of posts. For example, in Figure 1, the information recorded in post *a* is diffused from blog *A* to blogs *B* and *C*, and then from blog *B* to blog *E*.

The records of information diffusion that occur within the blog world are stored in the form of *diffusion history*. The diffusion history stores the time when a trackback or a scrap occurred, two bloggers who participated in the trackback or the scrap, and content on the post. By analyzing this information diffusion history, the phenomena of information diffusion occurring in the blog world can largely be divided into two types.

The first type is the information diffusion that occurs through explicit relationships in the blog world. In a general social network, information diffusion occurs through the word-of-mouth effect between members with a priori relationship and information spreads gradually among them. In the blog world, this corresponds to the phenomenon that a blogger trackbacks or scraps a post from one of the bloggers in her blogroll and posts are gradually disseminated among bloggers with such explicit relationships.

The second type is the information diffusion between blogs without explicit relationships. A blog service provider often selects some posts of quality contents and puts them on the main page of the blog world, which enables those who do not have any prior relationship with the owner of the post to be able to access it. In our preliminary experiments, it has been shown that about 85% of information diffusion within

the blog world occurs through this way. In this paper, when the information is diffused between blogs that do not have explicit relationships, we call it the information diffusion through *non-explicit relationship*.

With the information diffusion that occurs through explicit relationships, the number of blogs participating in the diffusion generally shows a tendency to increase linearly. However, with the information diffusion that occurs through non-explicit relationships, the number of blogs participating in the diffusion tends to increase exponentially. Figure 2 shows an example of this diffusion tendency with two sampled posts, where the horizontal-axis represents time and the vertical-axis represents the accumulated information diffusion power. Here, *information diffusion power* of a post is defined as the total number of blogs who directly or indirectly trackback or scrap the post in question. As time passes, the power of information diffusion occurring from non-explicit relationships tend to increase explosively, over 300 times the power of information diffusion arising from explicit relationships.

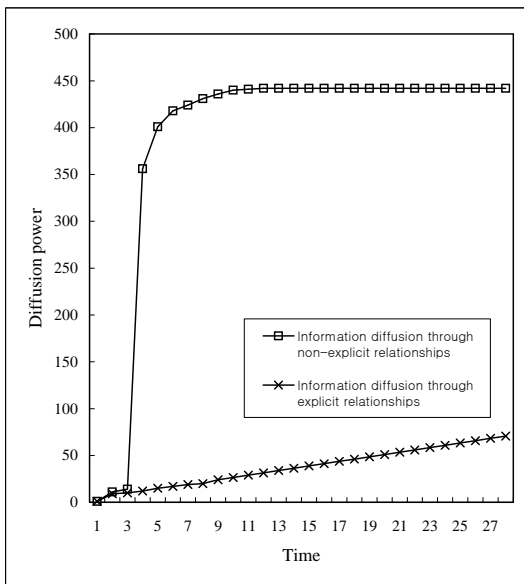


Figure 2: An example of explosive information diffusion.

When information spreads well in the blog world, it can be seen as bloggers actively using the blogs [31, 32]. In addition, if the patterns through which information is disseminated are discovered, a path that can spread information can be provided to bloggers. Therefore, it is important to model the phenomenon of information diffusion correctly in the blog world. In this paper, we address an information diffusion model that can explain well the phenomenon that information is diffused over the blog world through not only explicit relationships but also no-explicit relationships.

3. RELATED WORK

Existing studies on information diffusion in a social network have focused on modeling the word-of-mouth effect among social network members [12, 16, 18, 21]. The main idea shared by such studies is the following. Social network

members (i.e., nodes in a social network) may have influence on one another through their relationships (i.e., links in a network) and such influence may result in an influenced node who can influence other nodes in turn.

In [16], an independent cascade model was proposed to model information diffusion. The independent cascade model designates a probability to the relationship between nodes, and whether or not influence actually occurred is determined based on this probability. The independent cascade model simulates information diffusion phenomena and can be used to measure the number of nodes who are directly or indirectly influenced by the node in question. We call this value the *diffusion power* of the node in question.

In [18], a 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 value of the weighted influences received from her neighboring nodes is bigger than her threshold value, that node is regarded as having been influenced by the node who influenced her. Similar to the independent cascade model, the linear threshold model can be used to measure the diffusion power of a node.

In [21], a general cascade model was proposed. The general cascade model eliminates the condition in the independent cascade model that, in influencing a specific node, the neighboring nodes influence him independently, thus generalizing the characteristics of the linear threshold model and the independent cascade model. The general cascade model therefore is appropriate for explaining the diffusion phenomenon that shows the characteristics of both the linear threshold model and the independent cascade model.

In a blog world, a blogger trackbacks or scraps a post because she is influenced by the blogger who own that particular post. This can be modeled by the independent cascade model. Since trackback or scrap of a post is not a result of influences from many of her neighboring bloggers, the linear threshold model that calculates the diffusion power by adding up the weighted influences from one's neighbor nodes does not model well information diffusion in a blog world. Therefore, to explain information diffusion in a blog network, the independent cascade model is more appropriate than the linear threshold model or the general cascade model.

Various studies were carried out to analyze the phenomenon of information diffusion within a blog network [4, 11, 15, 24]. These existing studies established explicit relationships between blogs only when diffusion records between blogs existed during the analysis period, and explained information diffusion by analyzing the word-of-mouth effect through these relationships. In [11], the blog network information was classified according to themes and their diffusion tendencies were analyzed. In [4], weights were assigned to explicit relationships and, through them, a way to predict diffusion power and diffusion paths was proposed.

By establishing the relationship only when diffusion occurred, they carried out their analysis while essentially excluding the possibility of information diffusion between two blogs with no prior relationship in the future. Bloggers, however, can disseminate information from blogs with which they had no diffusion history during the analysis period, and such phenomenon frequently occurs in real world. This paper proposes a model that can explain the phenomenon of information diffusion through both explicit relationships and

paths other than explicit relationships.

4. PROPOSED MODEL

In this section, we propose a new model that considers both explicit and non-explicit relationships between blogs in order to explain all information diffusion phenomena in a blog world. We add new elements to an existing model that considers only explicit relationships between blogs so that the new model can explain information diffusion through non-explicit relationships as well. We also discuss ways to improve the accuracy of the proposed model and propose an expanded model.

4.1 Basic Information Diffusion Model

To explain information diffusion through explicit relationships, our basic information diffusion model uses an existing model. Here, an existing model refers to a technique of establishing explicit relationships between blogs, composing a blog network using those explicit relationships between blogs, and then analyzing information diffusion using the independent cascade model.

In order to explain information diffusion through non-explicit relationships, on the other hand, we add new elements of a super node, broadcast edges, and register edges to the existing model. Here, the super node represents the main webpage of a blog service provider and has bidirectional relationships with all blogs within the blog network. In this relationship, the link that goes from the super node to the blogs is called a *broadcast edge*, and the link that goes from the blogs to the super node is called a *register edge*. The super node can influence blogs through broadcast edges (when bloggers trackback or scrap some posts on the main page), and can get influenced by blogs through register edges (when some post from that blog has been selected and posted on the main page). Figure 3 shows the proposed basic information diffusion model.

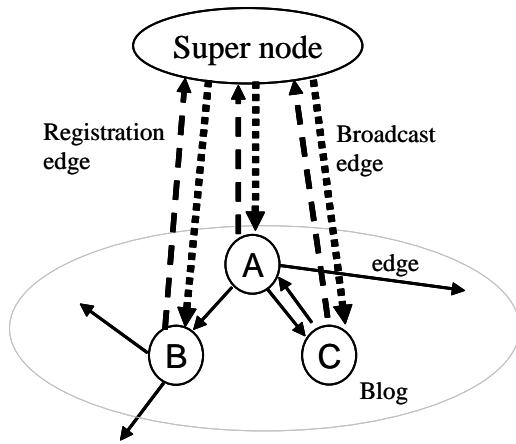


Figure 3: Basic information diffusion model.

Table 1 is a summary of terms and symbols needed to proceed with the discussion below. SN means the super node, U_i refers to user i . $P_{i \rightarrow j}$ is the probability that information on U_i will be disseminated to U_j , D_{SN} refers to the collection of posts posted on SN , and D_{SN} can possess m posts where m is the maximum number of posts that can be ex-

posed on SN . D_i refers to the collection of posts that U_i possesses, and $D_{i,j}$ refers to the j^{th} post of U_i . BE refers to the collection of broadcast edges. BE_i refers to broadcast edges directed from SN to U_i , and the assimilation probability of BE_i is shown as $P_{SN \rightarrow i}$. RE refers to the collection of register edges. RE_i refers to a register edge directed from U_i to SN , and the diffusion probability of RE_i is shown as $P_{i \rightarrow SN}$. If the total number of blogs within a blog network is n , BE and RE possess n edges each.

SN : super node
U_i : user i
$P_{i \rightarrow j}$: the probability that U_i information will be diffused to U_j
$D_i = \{D_{i,1}, D_{i,2}, \dots\}$: the collection of posts that U_i possesses
$D_{i,j}$: Document j of user i
$D_{SN} = \{D_{SN,1}, D_{SN,2}, \dots, D_{SN,m}\}$: the collection of posts registered on SN
BE_i : broadcast edge directed from SN to U_i
$BE = \{BE_1, BE_2, \dots, BE_n\}$: the collection of broadcast edges
RE_i : register edge directed from U_i to SN
$RE = \{RE_1, RE_2, \dots, RE_n\}$: the collection of register edges

Table 1: Summary of terms and symbols.

The diffusion probability of relationships existing within a blog network is calculated with the following method.

Equation (1) shows the computation of diffusion probability through *explicit* relationships between blogs. Bloggers write posts that may interest other users, but also those that may not interest others, such as posts about their private lives. Therefore, as the probability of U_i 's posts being diffused to U_j , the ratio of posts of D_i that other bloggers are interested in and those among them that are actually diffused to U_j is used. Here, in order to identify the posts among D_i , other bloggers are interested in, we identify only the posts which produced more than the threshold number θ of diffusions.

$$P_{i \rightarrow j} = \frac{\text{Among } D_j, \# \text{ of posts diffused from } D_i}{\text{Among } D_i, \# \text{ of posts having over } \theta \text{ of diffusions}} \quad (1)$$

Equation (2) shows the method of calculating the diffusion probability of *broadcast edges*. The probability of the super node diffusing U_i is $P_{SN \rightarrow i}$, and it is calculated as the ratio of posts actually diffused to U_i from D_{SN} .

$$P_{SN \rightarrow i} = \frac{\text{Among } D_i, \# \text{ of posts diffused from } D_{SN}}{|D_{SN}|} \quad (2)$$

Equation (3) shows the method of calculating the diffusion probability of *register edges*. The probability of a post by U_i diffusing the super node is $P_{i \rightarrow SN}$, and it is calculated as the ratio of posts among D_i that other bloggers are interested in and, among those, the posts that are actually posted on the

super node. As in equation (1), in order to identify posts among D_i that other bloggers are interested in, its diffusion history is analyzed, and only those posts that produced more than the threshold number θ of diffusions are used.

$$P_{i \rightarrow SN} = \frac{\text{Among } D_{SN}, \# \text{ of posts diffused from } D_i}{\text{Among } D_i, \# \text{ of posts having over } \theta \text{ of diffusions}} \quad (3)$$

Using the basic information diffusion model, we model information diffusion as follows: 1) Establish a blog network composed of explicit relationships between blogs; 2) Add to the blog network the super node, broadcast edges, and register edges; 3) Assign a diffusion probability to every relationship in the blog network using equations (1), (2), and (3); and 4) Analyze information diffusion by applying the independent cascade model to the blog network.

4.2 Expanded Information Diffusion Model

In this section, we propose an expanded information diffusion model that improves modeling accuracy of our basic model.

Bloggers may show various degrees of interest regarding the posts posted on the super node. Some posts may draw a lot of attention and be disseminated profusely while some other posts can be ignored by most bloggers. Figure 4 shows the distribution of diffusion power of posts actually posted on the super node. The horizontal-axis shows the ranking of posts based on their diffusion power, and the vertical-axis shows the diffusion power of posts in the number of trackbacks or scraps. Through Figure 4, it can be seen that the diffusion power of the posts on D_{SN} follows a *power function*.

In the basic information diffusion model, the differences in the diffusion power of posts are not considered in calculating the diffusion probability from the super node to a blogger; only a single diffusion probability $P_{SN \rightarrow i}$ is assigned to broadcast edge BE_i . Since posts with strong diffusion power and those with weak diffusion power are given the same diffusion probability, the accuracy in modeling information diffusion may have suffered.

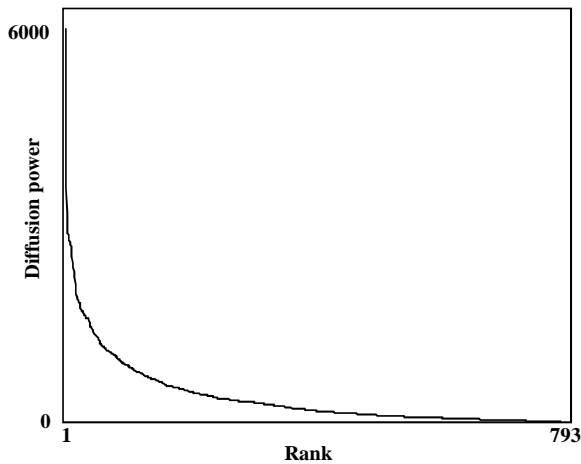


Figure 4: Distribution of diffusion powers of posts in the super node.

The new expanded diffusion model overcomes such a prob-

lem. Among the posts on D_{SN} , ones that have been composed by the same blogger are assumed to have the same diffusion power. Equation (4) shows the method of calculating probability $P_{i \rightarrow SN \rightarrow j}$ of the posts included in D_{SN} which were composed by U_i being diffused to U_j . $P_{i \rightarrow SN \rightarrow j}$ is calculated by multiplying α_i to $P_{SN \rightarrow j}$. Here, α_i is the correction coefficient of diffusion probability of the posts on D_{SN} that U_i composed.

$$P_{i \rightarrow SN \rightarrow j} = \alpha_i \times P_{SN \rightarrow j} \quad (4)$$

Equation (5) shows the method of calculating the correction coefficient of diffusion probability α_i regarding U_i . To calculate α_i , the average diffusion power of posts included in D_{SN} that were composed by U_i is divided by the average diffusion power of all the posts in D_{SN} .

$$\alpha_i = \frac{\text{Average diffusion power of } (D_i \cap D_{SN})}{\text{Average diffusion power of } D_{SN}} \quad (5)$$

The expanded information diffusion model is used as follows: 1) Establish a blog network using the basic information diffusion model; 2) Calculate the correction coefficient of diffusion probability for the composers of the posts in the super node, and compute the diffusion probability of those posts that reflect the correction coefficient; and 3) Analyze information diffusion by applying the independent cascade model to the blog network.

5. PERFORMANCE ANALYSIS

5.1 Experimental Setup

For experimental analysis, this paper used anonymized data which was collected from blog.naver.com, one of the largest blogospheres in Korea, for several months starting from July 2006. The number of posts created during the analysis period was about 100 million. For explicit relationships between blogs, *neighbor* relationships via the *blogroll* function were used. The main webpage of the blog service provider exposed 15 posts selected and changed them every 24 hours. During the analysis period, a total of about 1,200 posts were selected and exposed to bloggers.

For performance evaluation, we compared three methods: the original independent cascade model (ORIG) [16], the basic information diffusion model (BID) proposed in this paper, and the expanded information diffusion model (EID) which improved the basic information diffusion model.

As a performance metric, we used *error ratio*. First, we computed an *error*, the difference between the actual information diffusion history and the simulated based on each of three different models. Then, as a metric, we used the *error ratio* that compared the error values of two different diffusion models.

Equation (6) shows the computation of error ratio. M represents the diffusion model used, DH_i represents the actual history of information diffusion that is originated from blogger U_i . $Simulation(M; U_i)$ represents the information diffusion history from U_i , simulated by model M . Both the simulated and actual information diffusion histories are all represented as time series data. An element value of time series data indicates the number of blogs, accumulated everyday from the beginning of the whole analysis period, who

directly or indirectly trackback or scrap the posts of U_i . In our analysis, we need to compare the overall trends of information diffusion rather than its everyday difference. Therefore, we employed dynamic time warping (DTW) [9, 20, 22] to measure the difference (distance) between two time series data.

$$Error(M; U_i) = DTW(DH_i; Simulation(M; U_i)) \quad (6)$$

Equation (7) shows the method for computing the *error ratio*. $M1$ and $M2$ represent two different information diffusion models. The error ratio measures the ratio of the errors caused by two models while simulating information diffusion originated from U_i . The values between 0 and 1 mean that $M1$ produces less error than $M2$, while the values over 1 represent that $M1$ produces more error than $M2$.

$$ErrorRatio(M1; M2; U_i) = \frac{Error(M1; U_i)}{Error(M2; U_i)} \quad (7)$$

5.2 Experimental Result

In this section, three sets of experiments were analyzed. The first set of experiments compared the performance of the proposed basic information diffusion model (BID) and the original information diffusion model (ORIG). The second set of experiments compared the performance of the expanded information diffusion model (EID) and the original model (ORIG). The third set of experiments analyzed the performance of all three models together.

In the first set of experiments, we analyzed the error ratio of the basic information diffusion model over the original information diffusion model. Figure 5 shows the experimental results. Each point in Figure 5 depicts a post in the super node. The horizontal axis shows the diffusion power of each post, and the vertical axis does the error ratio of BID over ORIG when information diffusion of each post was simulated. The results show that the average error ratio of BID over ORIG is 1.65, which means that the error actually increased (65% more) when using BID than using ORIG. The reason for this is that BID does not reflect the varying degrees of diffusion powers of the posts in the super node.

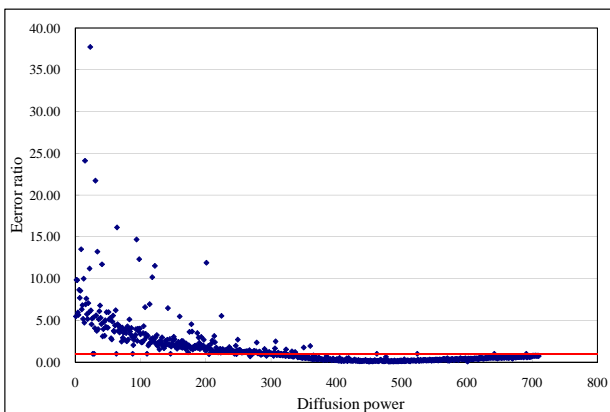


Figure 5: Error ratio (BID vs. ORIG).

We performed an additional experiment to test the performance of BID when the diffusion powers of posts in the

super node are similar. We selected a set of posts with diffusion powers similar to the average diffusion power of the posts in the super node, and measured the error ratio of BID over ORIG. The average diffusion power of the posts in the super node was 402 (that is, the number of bloggers who directly or indirectly trackback or scrap the post in question is 402 on average), and we sampled 44 posts whose diffusion power is within the range of [350, 450].

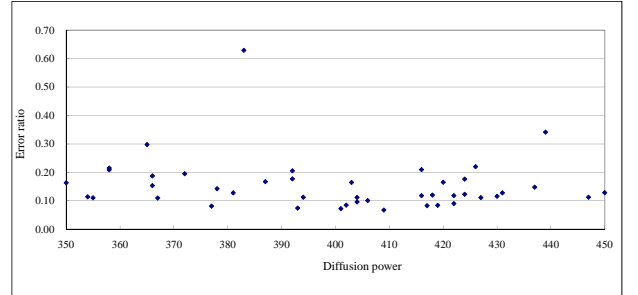


Figure 6: Error ratio of posts with average diffusion power (BID vs. ORIG).

Figure 6 shows the results. Each point in Figure 6 signifies a post with the diffusion power between [350, 450]. As in the first experiment, the horizontal axis shows the diffusion power of posts, and the vertical axis does the error ratio of BID over ORIG. The error ratio is 0.15, which means the error occur 85% less in BID than ORIG. This confirms our observation that BID successfully models information diffusion of the posts in the super node as long as the diffusion power of posts are similar to each other, which results in the improvement model of BID, the EID.

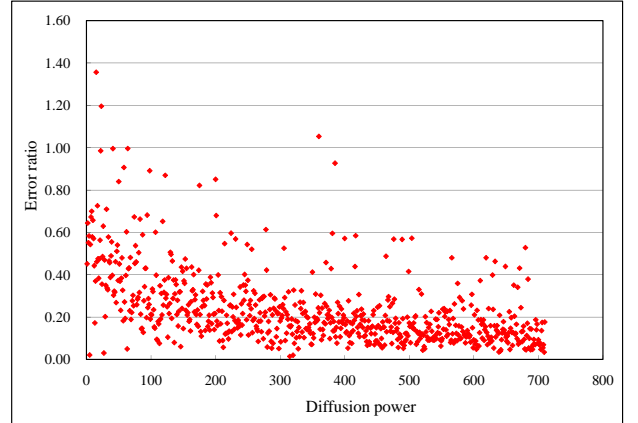


Figure 7: Error ratio (EID vs. ORIG).

In the second set of experiments, we analyzed the error ratio of EID and ORIG. Figure 7 shows the experimental results. Each point in Figure 7 signifies a post in the super node. The horizontal axis shows the diffusion power of each post, and the vertical axis shows the error ratio of EID over ORIG. The results show that the average error ratio of EID over ORIG is 0.23, which means that errors occur 77% less when using EID than using ORIG. Different from the previous result with BID (in Figure 5), EID explains the information diffusion phenomenon very well even when there

Group Statistics

	EID	N	Mean	Std. Deviation	Std. Error Mean
BID	1.00	710	0.7660	0.17401	0.00653
	2.00	710	-0.6547	2.75570	0.10342

Independent Samples Test

		Levene's Test for Equality of Variances		T-test for equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
BID	Equal variances assumed	271.843	0.000	13.710	1418	0.000	1.42073	0.10363	1.21745	1.62400
	Equal variances not assumed			13.710	714.654	0.000	1.42073	0.10363	1.21728	1.62417

Figure 9: Performance comparison between EID and ORIG based on T-test.

exist the posts with varying degrees of diffusion powers in the super node.

In the third set of experiments, we compared the performance of ORIG, BID, and EID with the real diffusion history, DH. Figure 8 shows examples of diffusion history of 4 sampled documents. Each graph shows the diffusion history of a post, either simulated or actual. Note that each graph corresponds to a single point in Figures 5 to 7. The horizontal axis represents time, and the vertical axis represents the diffusion power of a particular post. The experimental results show that EID is the most similar to DH, followed by BID and then ORIG. Since ORIG takes into account information diffusion through explicit relationships only, the diffusion power of a post increases linearly, while BID and EID models the explosive information diffusion, which is also observed in DH.

Finally, Figure 9 also verifies the superior performance of EID over BID based on T-test.

6. CONCLUSIONS

Modeling information diffusion in a blog world is a useful research venue, which can be used for predicting information diffusion, detecting abnormality, marketing, and revitalizing the blog world. Existing studies established explicit relationships between blogs, and analyzed only the word-of-mouth-effect through such explicit relationships. However, more than 85% of all information diffusion in a blog world happens through non-explicit relationships, such as information diffusion through a main webpage of a blog service company.

This paper has proposed a basic model that explains both the phenomena of information diffusion through explicit relationships and those through non-explicit relationships. The new model adds a super node, broadcast edges, and register edges to the existing information diffusion model. This paper has also proposed a method to assign the assimilation probability to each relationship.

Since the basic model does not take into account that the

posts in the super node may have varying degrees of diffusion powers, the performance may not be as good. In this paper, we have proposed an expanded information diffusion model that improves the performance of the basic model by taking into account the degrees of diffusion powers of posts in the super node.

Through experiments, we have verified the new models are better than the existing information diffusion model. The experimental results reveal that, when the diffusion powers of posts are similar to each other, the basic information diffusion model generates 85% less errors than the existing model. Even when the diffusion powers of posts vary significantly, the expanded information diffusion model generates 77% less errors than the existing model.

In this paper, we have considered only the main webpage of a blog service company as the source of information diffusion through non-explicit relationships. However, most blog service companies provide search engine services for finding information in the blog world, which is another cause of information diffusion through non-explicit relationships. At present, we are investigating the ways to incorporate the search engines into the expanded information diffusion model.

7. ACKNOWLEDGMENT

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. Also, this work was supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea government(MEST)(No. R01-2008-000-20872-0).

8. REFERENCES

- [1] B. Aaron et al., "Equating R-Based and D-Based Effect-Size Indices: Problems with a Commonly Recommended Formula," *Florida Educational Research Association*, 1998.

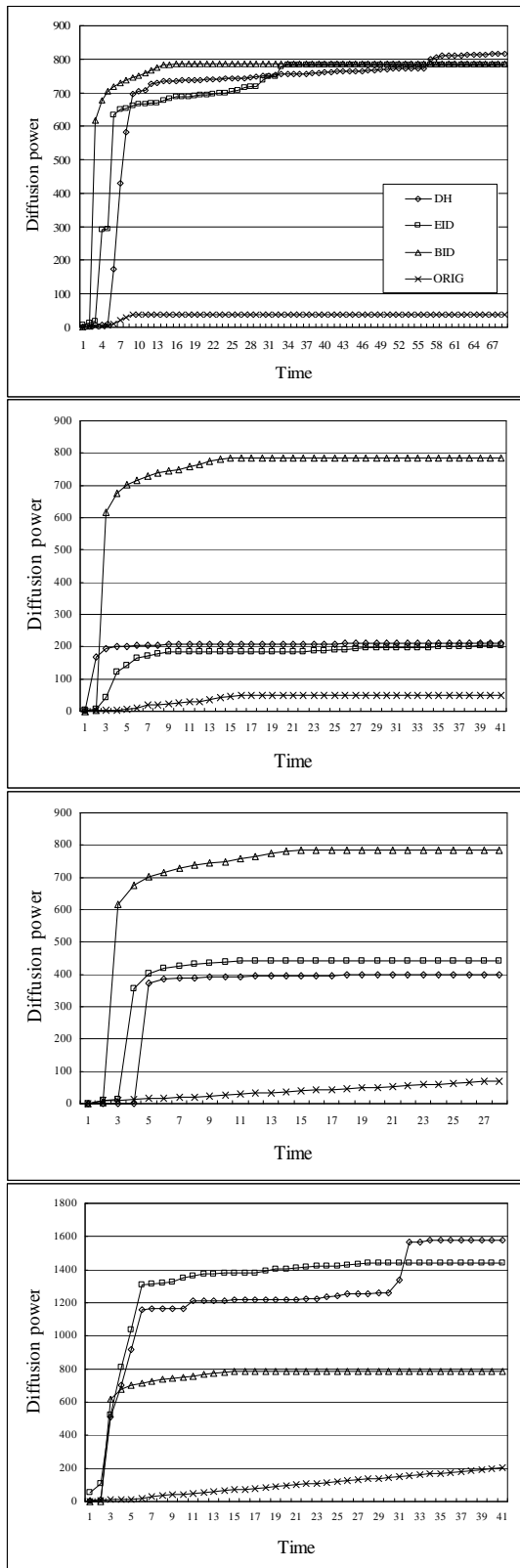


Figure 8: Simulation of information diffusion.

[2] L. Adamic, O. Buyukkokten, and E. Adar, "A Social Network Caught in the Web," *First Monday*, Vol. 8, No.

6, pp. 1-22, 2003.

[3] N. Agarwal et al., "Identifying the influential bloggers in a community," In *Proc. Int'l. Conf. on Web Search and Web Data Mining*, WSDM, pp. 207-218, 2008.

[4] A. Java et al., Modeling the Spread of Influence on the Blogosphere, Technical Report TR-CS-06-03, University of Maryland, Baltimore, 2006.

[5] C. Asavathiratham et al., "The Influence Model," In *Proc. IEEE Int'l. Conf. on Control Systems*, pp. 52-64, 2001.

[6] R. Albert, H. Jeong, and A. Barabasi, "Diameter of the World Wide Web," *Nature*, Vol. 47, pp. 651-654, 2000.

[7] Blogger.com Co., Ltd. <http://blogger.com>

[8] J. Brown and P. Reinegen, "Social Ties and Word-of-Mouth Referral Behavior," *Journal of Consumer Research*, Vol. 1, No. 3, pp. 350-362, 1987.

[9] C. Ratanamahatana, and E. Keogh, "Making Time-Series Classification More Accurate Using Learned Constraints," In *Proc. SIAM Int'l. Conf. on Data Mining SDM 2004*.

[10] SK Communications, <http://www.cyworld.com>

[11] D. Gruhl et al., "Information Diffusion Through Blogspace," In *Proc. Int'l. Conf. on World Wide Web, WWW*, pp. 491-501, 2004.

[12] P. Domingos and M. Richardson, "Mining the Network Value of Customers," In *Proc. ACM Int'l. Conf. on Knowledge Discovery and Data Mining, ACM SIGKDD*, pp. 57-66, 2001.

[13] G. Ellison, "Learning, Local Interaction, and Coordination," *Econometrica*, Vol. 61, No. 5, pp. 1047-1071, 1993.

[14] Empas Corp., <http://www.empas.com>

[15] F. Duarte et al., "Traffic Characteristics and Communication Patterns in Blogosphere," In *Proc. Int'l. Conf. on Weblogs and Social Media, ICWSAM*, 2007.

[16] 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*, Vol. 12, No. 3, pp. 211-223, 2001.

[17] M. Granovetter, "The Strength of Weak Ties," *American Journal of Sociology*, Vol. 78, No. 6, pp. 1360-1380, 1973.

[18] M. Granovetter, "Threshold Models of Collective Behavior," *American Journal of Sociology*, Vol. 86, No. 6, pp. 1420-1443, 1978.

[19] iSAVEZONE Corp., <http://www.isavezone.com>

[20] E. Keogh, "Exact Indexing of Dynamic Time Warping," In *Proc. Int'l. Conf. on Very Large Data Bases, VLDB*, pp. 406-417, 2002.

[21] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the Spread of Influence through a Social Network," In *Proc. ACM Int'l. Conf. on Knowledge Discovery and Data Mining, ACM SIGKDD*, pp. 137-146, 2003.

[22] S. Kim, S. Park, and W. Chu, "An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases," In *Proc. IEEE Int'l. Conf. on Data Engineering, IEEE ICDE*, pp. 607-614, 2001.

[23] R. Kumar, J. Novak, and A. Tomkins, "Structure and Evolution of Online Social Networks," In *Proc. Int'l. Conf. on Knowledge Discovery and Data*, pp. 611-617,

2006.

- [24] M. McGlohon et al., "Finding Patterns in Blog Shapes and Blog Evolution," In *Proc. Int'l. Conf. on Weblogs and Social Media*, 2007.
- [25] S. Milgram, "The Small World Problem," *Physiology Today*, Vol. 2, pp. 60-67, 1967.
- [26] MySpace.com Co., Ltd. <http://www.myspace.com>
- [27] NHN Corp., <http://www.naver.com>
- [28] A. Nowak, *Virus Dynamics: Mathematical Principles of Immunology and Virology*, Oxford University Press, 2000.
- [29] S. Redner, "How Popoular Is Your Paper?," *European Physics Journal B*, Vol. 4, No. 2, pp. 131-134, 1998.
- [30] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994.
- [31] D. Watt and S. Strogatz, "Collective Dynamics of 'Small-World' Networks," *Nature*, Vol. 393, pp. 440-442, 1998.
- [32] D. Watts, *Small Worlds: The Dynamics of Networks Between Order and Randomness*, Princeton, New Jersey: Princeton University Press, 1999.