An Analysis of Information Diffusion in the Blog World

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

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

Sunju Park
School of Business
Yonsei University, Korea
boxenju@yonsei.ac.kr

ABSTRACT

In the blog world, bloggers produce information, establish relationships with other bloggers in order to exchange information, and form a blog network, an online social network. In social network theory, information diffusion is said to occur through the relationships of the social network's constituents. Contrary to the social network theory, however, when the history of information diffusion that occurs in the actual blog world is examined, a majority of information spreads between the constituents without any established relationship. Moreover the phenomenon of explosive information diffusion is also observed. Using the real-world blog data, this paper reveals that these two phenomena are related to each other and examines the causes that induce explosive information diffusion.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: SOCIAL AND BEHAVIORAL SCIENCES—*Sociology*

General Terms

Human Factors, Algorithms, Experimentation, Economics

Keywords

Social Network Analysis, Blog, Data Mining, Information Diffusion

1. INTRODUCTION

A social network captures the social structure among members of a society (either individuals or organizations) through a network composed of nodes and edges that represent members and their relations, respectively.

The blog world is an online social network[2, 4] composed of bloggers [3, 5, 7, 8, 9, 10, 11, 13, 14, 15]. In the blog world, a function called *blogroll* enables bloggers to enter into a relationship with other bloggers either to obtain needed information easily or to maintain a friendly relationship. This is similar to the shortcut function of web browsers, such as favorites lists of bookmarks, and it helps one to easily visit the blogs or his/her interest at anytime. In this paper, a relationship formed between two blogs through blogroll is defined as an *explicit relationship*, and an online social network composed of explicit relationships is defined as a *blog network*. Note that other types of explicit relationships, such as bookmarks, messangers, and texting, may exist between blogs, but since those are not observable, we only consider relationships between blogs through blogroll as explicit relationships.

The 'trackback' function allows bloggers to make their posts linked to other bloggers' posts [5, 14]. Also, the 'scrap' function, provideded by most blog-service companies in Korea [7, 10, 15], allows bloggers to copy other bloggers' posts to their blogs. The act of a blogger's trackbacking or scrapping other blogger's post causes the information in that post to spread. A post becomes trackbacked or scrapped to other blogs, which again becomes trackbacked or scrapped to yet more blogs, and through such repetitions, the information is diffused through the blog world. When this information diffusion in the blog world is analyzed, the information diffusion paths can be found. Such analysis can be used to measure the activity level of the blog world and also to predict the diffusion pattern of newly created information.

In social network theory, information diffusion in the social network is said to occur through the established relations between members[6]. In the real world, explicit relationships are typically established through direct contact between acquaintances. When applying the social network theory to the blog network, that means the information diffusion shall occur through explicit relationships established between blogs. Our examination of information diffusion history revealed that, contrary to the existing theory, a majority of information occurred between blogs with no explicit relationships. Also, some posts show an explosive increase in the number of blogs who trackback or scrap those posts. This paper examines the information diffusion in the blog world with respect to these two phenomena.

2. BLOG WORLD

The blog world is an online social network composed of blogs [5, 7, 10, 14, 15]. A *blog* is a type of a personal website where personal thoughts or opinions can be recorded as *posts*. 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

CNIKM'09, November 6, 2009, Hong Kong, China.

can be done on someone else's blogs, such as reading the posts of others, leaving one's own opinions as comments, scrapping (copying) posts to one's own blog, or trackbacking by leaving a link, which leads to one's own blog where one's opinions about the same topic have been written and recorded. Through these actions regarding posts, a blogger can discuss his or her thoughts with other bloggers. In the blog world, similar to the hyperlink of the web, a function called *blogroll* is used to establish an explicit relationship with another blog that the blogger is interested in. Because these established relationships are all preserved, a blogger can easily move to any of the relevant blogs at any time through those relationships.

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 relationships between blogs. The dotted 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 trackbacked by blogs *B* and *C* and trackbacked to blog *E* in turn.

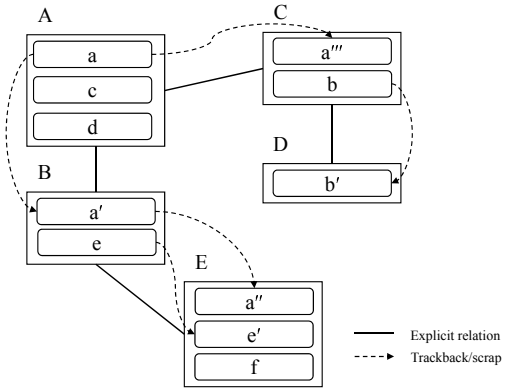

Figure 1: Blog world example.

3. INFORMATION DIFFUSION WITHOUT EXPLICIT RELATIONSHIP

All information diffusion that occurs within the blog world is stored as diffusion history. Information diffusion history records such information as the times when trackbacking or scrapping occurs, bloggers involved, posts, etc. in the form of logs. Also, all explicit relationships established between bloggers are recorded. When the information diffusion history and the explicit relationships between blogs are compared and examined, two phenomena which are contrary to the existing social network theory are discovered: (1) a majority of information diffusion takes place between blogs without any explicit relationship, and (2) the explosive information diffusion is observed in some cases.

First, the information seems to spread between blogs without any explicit relationship. According to the existing social network theory, information diffusion within a social network is said to generally occur through the explicit relationships between the members. However, when the information diffusion history in a blog network is examined, many of the information diffusion between blogs with no explicit relationship take place. That is, even when no explicit relationship exists between the two blogs, trackbacking or scrapping has occurred. We define the ratio of the diffusion between blogs without explicit relationships during the observation period as follows.

$$R = \frac{\text{number of diffusion without explicit relationships}}{\text{number of total diffusion}}$$

Using the real-world data, we examine the ratio of diffusion without explicit relationships. Analysis was carried out on 10,000,000 original posts produced during the analysis period at one of the most popular blog site in Korea. When the ratio of diffusion without explicit relationships is examined, the results are as follows. During the analysis period, a total of 26,190,503 trackbacks or scraps occurred, and among them, a total of 22,281,385 trackbacks and scraps occurred between blogs without explicit relationships. That is, diffusion between blogs without explicit relationships takes up 85% of the total information diffusion.

Second, the phenomenon of explosive information diffusion is observed. The number of blogs who trackbacks or scraps the information in a post is called the *information diffusion degree*. The diffusion degree of various posts reveals that in most cases, the information diffusion degree increases linearly as time passes by. With some posts, however, their diffusion degree increases explosively from a certain point on.

Figure 2 shows an example of the explosive information diffusion. The graph shows the changes of information diffusion degree of two posts through time. It can be seen that the information diffusion degree of the post drawn with solid line increases linearly as time passes by. The post drawn with dotted line, however, shows explosive information diffusion. From a certain point at time 3, it shows an increase of information diffusion degree that averages over 300 times of the regular information diffusion degree.

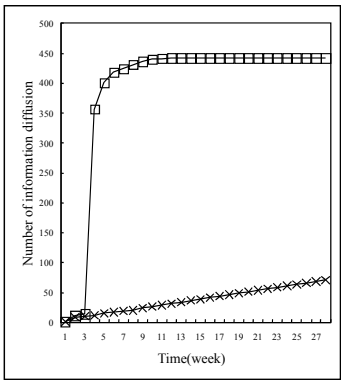

Figure 2: Explosive increase in information diffusion.

A close look at all the posts which show such explosive information diffusion degree reveals that the information diffusion between blogs without explicit relationships also occurs with them. It seems that the explosive information diffusion and the diffusion without explicit relationships were occurring from the same cause. In the following, the causes that induce the diffusion without explicit relationships and explosive information diffusion are carefully examined using data mining techniques. The results of this examination can be

useful in many applications related to blog network research, such as predicting the potential of explosive information diffusion and the information diffusion degree.

Table 1: Potential attributes that can cause explosive information diffusion.

• use of search engines (allowing posts to be exposed to search engines)
• being listed on a portal main page
• being listed on a blog portal main page

4. CAUSES FOR EXPLOSIVE INFORMATION DIFFUSION

Table 1 shows the potential attributes of a post that can cause explosive information diffusion. The use of a search engine is an attribute for whether or not a blogger allows his/her posts to be searched by search engines. When searching is allowed, other blogs without explicit relationships can use the search engine to visit one's blog and trackback/scrap the posts. Being listed on a portal main page is an attribute of whether or not one's posts have been exposed on the main page of the portal website. The portal selects a small number of posts from the blog world and exposes them on its main page. The blog portal main page is similar to the portal main page and refers to a separate main page specifically for bloggers.

To discover which attributes are connected to the explosive information diffusion, we analyze the correlation between the group that shows explosive information diffusion and the potential attributes in Table 1 is analyzed.

First, in order to deduce the groups that show explosive information diffusion, all posts that are created during the analysis period are clustered based on their diffusion degree. The posts with similar diffusion degrees can be judged to have the same diffusion tendency, and each of the formed clusters signifies an information diffusion tendency. After clustering, the correlation between the formed clusters and potential attributes is identified by analyzing the characteristics of every group. Through this process, it can be known whether or not explosive information diffusion is one of the tendencies of information diffusion and what attributes cause such tendency. The above clustering process and the process of analyzing characteristics are carried out together to deduce a more reliable correlation between clusters and attributes [12].

Table 2 lists the causes for explosive information diffusion. *Cluster number* is the identifier of a derived cluster. *Cause ex* signifies diffusion through explicit relationship, while *cause pm* signifies diffusion due to being listed on a portal main page. *Cause bm* signifies diffusion due to being listed on a blog portal main page, while *cause se* signifies diffusion through a search engine.

In Table 2, cluster 1 is connected with the two attributes *causes ex* and *se*. When the average, minimum, and maximum diffusion number are examined, all the posts in cluster 1 with *cause ex* have been scrapped only once. It can be seen that of the posts being diffused between blogs with explicit relationships, 78.2% belong to cluster 1. In other words, diffusion through explicit relationships is connected with the diffusion tendency of cluster 1. Also, when *cause se* is examined, it can be seen that among the posts being diffused through search engines, 39.2% belong to cluster 1. This, too, shows that diffusion through search engines is connected with the diffusion tendency of cluster 1. In the case of clusters 2 and 3, they also show similar tendencies as that of cluster 1.

When cluster 18 is examined, it shows the connection with the attribute of being listed on a portal main page. The posts in the cluster were trackbacked or scrapped on average 55.2 times, and the highest number was 3,961 times. When *cause pm* is examined, it can be seen that among the posts listed on a portal main page, 90.9% belong to cluster 18. Therefore, the diffusion of posts listed on a portal main page shows the connection to the diffusion tendency of cluster 18.

The characteristic of cluster 8 is that it is composed of posts that have been trackbacked or scrapped on average 13,005.64 times. The number is far above those of other clusters. Also, all posts that make up this cluster are the posts which are diffused through search engines. Accordingly, it can be seen that the diffusion of posts through search engines is connected with the diffusion tendency of cluster 8.

The result of analysis shows that no cluster was derived which was connected to what would be identified as *cause bm*, namely, being listed on a blog portal main page. In other words, it was shown that being listed on a blog portal main page was not a sufficient cause that induces information diffusion tendency.

Table 3 shows the verification of correlation between potential causes connected to each of the derived clusters to within 0.005% of significance level using the chi-square method [1]. According to the table, the correlation between each cluster and corresponding causes is statistically significant.

Table 4 divides the posts according to each potential causes and shows their characteristics. As shown in Table 4, the number of posts which have been listed on a portal main page is rather small, but their average diffusion per post is 385.78 times, and it is especially contrastive when compared to the diffusion through explicit relationships, the average of which is 1.57 times.

Table 3: Statistical significance level of correlation between clusters and potential causes(significance level 0.005%).

Cluster No	Causes	χ^2	Sig.
1	explicit(ex)	7,165,881	ok
1	search engine (se)	3,388,558	ok
2	explicit(ex)	1,403,548	ok
3	explicit(ex)	573,073	ok
18	potal main (pm)	472,476	ok
8	search engine (se)	23,088	ok

Table 4: Statistics of each potential cause.

Causes	Number of posts	Number of diffusion		
		Total	Avg.	Max.
explicit(ex)	3,909,145	2,486,568	1.57	3,878
potal main (pm)	463,218	1,207	383.78	5,862
search engine (se)	21,818,123	8,505,123	2.56	16,521
Total	26,190,503	9,971,149	2.62	16,521

Table 2: Correlation between potential causes and deduced information diffusion clusters.

Cluster No	Cause	Number of Posts	Number of posts for each cause			Number of diffusion		
			0	1	3	Avg.	Min.	Max.
1	ex	7,436,021	1,944,011(78.2%)	13(1.1%)	5,491,997(39.2%)	1	1	1
1	se	7,436,021	1,944,011(78.2%)	13(1.1%)	5,491,997(39.2%)	1	1	1
2	ex	1,704,028	311,665(12.5%)	7(0.6%)	1,392,356(9.9%)	2	2	2
3	ex	679,015	106,074(4.3%)	5(0.4%)	572,936(4.1%)	3	3	3
18	pm	462,885	11,727(0.5%)	1,097(90.9%)	450,061(3.2%)	55.52	29	3,961
8	se	747,214	0(0%)	0(0%)	747,214(5.3%)	13,005.64	12,451	13,585

Information diffusion between blogs without explicit relationship occupies 85% of the total information diffusion. The analysis of the correlation between potential causes that can induce explosive information diffusion and various tendencies of information diffusion reveals two causes. They are being listed on the portal main page and being diffused through search engines. These are also the attributes that cause diffusion between blogs without explicit relationships.

The attribute of being listed on a portal main page shows the following characteristics. First, the posts listed on a portal main page are exposed to all bloggers in the blog world. When people log on to their blogs in order to participate in the blog world, all bloggers view their portal main page. Accordingly, the posts that are exposed on the main page tend to show an explosive diffusion tendency. Second, the posts exposed on a portal main page are manually selected by the blog-service provider. Third, the portal main page has space and time limitations.

The attribute of being diffused through search engines shows the following characteristics. First, information diffusion occurs based on the interest of the blogger, not based on the relationships between bloggers. Because the blogger is searching the search engines for information that he or she is interested in, diffusion occurs based on information content, not based on the relations in a blog network. Second, the information diffusion degree varies significantly. This is because both information diffusion which reflects the personality of each blogger and information diffusion which reflects the general interest of the society occur such as current hot issues through search engines.

5. CONCLUSIONS

According to the social network theory, information diffusion in the blog world should occur through explicit relationships between blogs. When the information diffusion history was analyzed, however, information diffusion between blogs without explicit relationships was discovered. In addition, it was found that through this information diffusion, in time, explosive information diffusion could occur. The analysis showed that the diffusion between blogs without explicit relationships occupied 85% of the total volume of diffusion in the blog world. That is, the diffusion between blogs without explicit relationships is an important information-diffusion tendency of the blog world. By using the clustering and characteristics analysis, we have discovered two causes of explosive information diffusion are (1) being listed on a blog world service portal main page and (2) being diffused through search engines.

This paper explains that the phenomenon of an online social network of blogs is different from what is generally known to be true in social network theory, and confirms that this phenomenon occupies an important part in the blog world's information diffusion.

6. ACKNOWLEDGMENT

This work was partially supported by the Korea Science and Engineering Foundation(KOSEF) grant funded by the Korea goverment(MEST)(No. R01-2008-000-20872-0) and the ITRC support program supervised by the IITA (IITA-2009-C1090-0902-0040).

7. REFERENCES

[1] B. Aaron et al., "Equating r-Based and d-Based Effect-size Indices: Problems with a Commomly Recommended Formula," *Florida Educational Research Association*, 1998.

[2] L. Admic, 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] R. Albert et al., "Diameter of the World Wide Web," *Nature*, Vol. 47, pp. 651-654, 2000.

[5] Blogger.com Co., Ltd. *http://blogger.com*

[6] 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.

[7] SK Communications, *http://www.cyworld.com*

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

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

[10] iSAVEZONE Corp., *http://www.isavezone.com*

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

[12] Y. S. Kwon and S. W. Kim, Clusters Characterization: A Data Mining Technique Performing Clustering and Characterization Simultaneously, Technical Report Hanyang University 2007.

[13] M. McGlohon et al., "Finding Patterns in Blog Shapes and Blog Evolution," In *Proc. Int'l Conf. on Weblogs and Social Media*, ICWSAM, 2007.

[14] MySpace.com Co., Ltd. http://www.myspace.com

[15] NHN Corp., *http://www.naver.com*