Determining the Strength of the Propensities of a Blog Network

Abstract

A blog network may exhibit two different propensities characterized by the purpose of use: an information-oriented propensity and a friendship-oriented propensity. Both propensities coexist in a blog network, and the degree of these propensities plays an important role in business and policy decisions of blog-related business. In this paper, we propose an automated method for measuring the propensities of a blog network. First, we use classification to judge the propensity values of a relation between two blogs. Then, by adding up the propensity values of all the relations in the network, we determine the propensity values of the whole network. Two normalization methods are proposed to solve the potential problem that the propensity value tends to increase with the increasing size of a network. Our approach is validated through extensive experiments using a large volume of real-world blog data. The experimental results show that our method achieves a high level of accuracy in determining the propensity values of a relation. The results also suggest the applicability of our approach to determining the propensity values of an entire network.

Keywords: blog networks, social network analysis, propensity determination

1. Introduction

A blog is a personal website where its owner (blogger) stores diaries, commentaries, and news of his/her interests [18][19][20][23]. The blogger publishes posts in his/her blog and establishes relations with other blogs through links, comments, trackbacks, blogrolls, etc. While used for expressing one's thoughts and opinions in early days, blogs have evolved for connection, interaction, and other social activities [19]. In this respect, a blogosphere can be viewed as an online social network composed of nodes that represent blogs and links that represent relations between blogs [4][7]. We call such an online social network a 'blog network.' Typical examples of blog networks include myspace.com, blogger.com, facebook.com, etc.

As 'blogging' became popular to a large population, companies have become interested in providing products and services that utilize blogs. To run a successful blog-related business, companies have started to analyze the blog network, investigating various properties of the blog network, such as the scale of a blog network, the age distribution of bloggers, etc. [18][19][20][23].

Among many properties of the blog network, one of the most important is the purpose of use, that is, why the blogger uses blogs [8][9]. The Web survey of 1,000 bloggers in Korea shows that the purposes of using blogs are documenting and sharing daily life, sharing experience, sharing fun, storing and sharing information, and sharing opinions [19]. The purpose of use can be largely divided into two categories: for managing information and for maintaining acquaintance.

In this paper, the different behavior pattern exhibited in a blog network according to the different purpose of use is defined as 'propensity.' We identify two propensities: Information-

oriented Propensity (IoP) and Friendship-oriented Propensity (FoP). The IoP is defined as the propensity of using the blog network to search, disseminate and share information, while the FoP is defined as the propensity of using the network to establish and encourage interpersonal interactions.

Blog-service providers and marketing companies should provide products and services appropriate for the purpose of use [2][16][17]. In a blog network where bloggers are mostly interested in information, for example, viral marketing which utilizes the phenomenon of information dissemination can be effective. By providing the bloggers with the services and functionalities for easy collection and dissemination of information, blog companies aid the spread of information through blogs. Marketing companies may also utilize blogs by creating informative contents about their products for promotion [20][22][23].

In a blog network where bloggers are mostly interested in interpersonal interactions, on the other hand, blog companies may entice bloggers to participate in a variety of group events and promotions. Also, blog-service companies may sell cyber-items such as skin or music that can be used as gifts among friends. Social applications, that have shown explosive growth in recent years, are the primary example of utilizing the FoP.

That is, different business strategies should be applied based on the degree of the propensities of a blog network. The propensity values of a network might be measured through direct investigation and analysis by domain experts. Using domain experts, however, is implausible for two reasons. First, the size of a network may be too large for a human expert to determine the propensity values. Second, even if the propensities can be measured, it can be very time-

consuming. A blog network tends to be very dynamic, and its propensities may change over time, which makes it hard to use human experts every time the propensity judgment is needed.

In this paper, we propose an automated method that can be used to determine the strength of the propensities of a blog network. The main idea is to compute the IoP and FoP values of a relation between two blogs, the basic unit of a network, and using them to determine the propensity values of the whole network.

The accuracy of the proposed method and its applicability are demonstrated through extensive experiments. The results show that our classification-based method achieves a high level of accuracy in evaluating the propensity values of a relation. To demonstrate the validity of our approach in determining the propensity values of a whole network, we compute and compare the propensity values of a blog network that seems to exhibit a particular propensity more strongly and those of a randomly-extracted blog network. The results show that our approach correctly identifies the network with strong propensity as such.

The rest of the paper is organized as follows. Section 2 defines the problem. Section 3 briefly describes related work. Section 4 explains in detail how to compute the propensity values of a relation and those of a network. Section 5 demonstrates the accuracy of the proposed method through experiments. Section 6 concludes the paper.

2. Problem Definition

In a blogosphere, there exists a wide variety of connections between its members, and through these connections, members form relations. Figure 1 depicts an example of a blogosphere. Each

rounded rectangle, labeled 1 through 11, represents a blog, whereas each small rectangle represents a post in a blog. The arrow represents a connection between two bloggers. A connection can be either from a blog to a blog (e.g., keeping a blogroll), from a blog to a post (e.g., putting comments on someone's post), from a post to a post (e.g., keeping a link or trackback), or from a post to a blog (e.g., keeping a link to another blog).

<Insert Figure 1 here.>

A social network is a social structure made of nodes that represent its members and relations that represent the connections between its members [3][21]. A blogosphere can be viewed as a social network, which we call a blog network. Depending on how nodes and relations are defined, a blogosphere can display different network topology. A node may represent a blog or a post. A relation may be defined as a connection from a post to another post and/or a connection from a blog to another blog. Figure 2 depicts a network diagram of the blogosphere in Figure 1, where blogs are represented as nodes, and all connections are represented as edges. The numbers in Figure 2 denote the blog numbers in Figure 1.

<Insert Figure 2 here.>

Having the bloggers who actively participate in blog activities and interact with others, the blog network shows a great potential for advertisement and communication. Active interaction and fast dissemination in the blog network also makes the blog world an efficient medium for word-of-mouth marketing [5][8][11]. Prior research suggests that word-of-mouth marketing plays an important role in customer's purchasing decisions. As the choice of products and services increases, customers become more reliant to the advice of the people around at the first phase of their purchasing process [20][23].

For advertisement and marketing in a blog network to be effective, one needs to take into account why bloggers are using blogs and how they interact with other bloggers, that is, the purpose of use. In this paper, we divide the behavior of bloggers into two propensities: Information-oriented Propensity (IoP) and Friendship-oriented Propensity (FoP). Bloggers may publish information-oriented articles in such topics as stocks, travels and news, and read and collect similar types of articles from other blogs. We define the propensity of the blog activities for information collection and dissemination as Information-oriented Propensity (IoP). Bloggers may post articles of more personal nature, such as personal news and diaries, to build, maintain, and develop personal relationship with others. We define the propensity of the blog activities for interpersonal interactions as Friendship-oriented Propensity (FoP).

Of course, both IoP and Fop coexist in a single network. The blogger displays different propensities depending on the blog he/she visits. That is, the blogger may exhibit either the IoP or the FoP, depending on his/her counterpart. Most bloggers display both propensities to a certain degree, and therefore most blog networks show both propensities.

In order to run a successful blog-related business, companies need to apply the business practice best suited for the target network. In particular, it is important to know the degree of the IoP and the FoP of the network. In this paper, we address the problem of determining the strength of the two propensities of a blog network.

3. Related Work

Little research exists that addresses the problem of determining the propensities of a blog network. In this section, we review previous research from the field of classification and data mining and discuss their applicability to our problem.

Agrawal *et al.* proposed a method to categorize the writers on a given topic in a newsgroup into two opposite classes: one is "for" and the other is "against" the topic [1]. Their idea is based on the observation that writers usually make quotations when they disagree than when they agree. Let's suppose post *A* is posted in a newsgroup. The writer of post *B* which quotes the original post *A* is regarded as to have an opinion opposite to the writer of post *A*. In a similar fashion, the writer of post *C* which quotes post *B* is regarded as to have an opinion opposite to the writer of post *B*. Since the writer of post *A* and the writer of post *C* both have an opinion opposite to the writer of post *B*, they are considered to have the same opinion. Quotations in newsgroup posts are used to infer an implicit social network among writers who participate in the newsgroup. Their graph-theoretic algorithm achieves high accuracy in classifying writers into two groups.

Their research and ours have it in common that both focus on classification. In our problem, however, the target to be classified is a blog network rather than individual blogs. Furthermore, our problem is not to classify a blog network into one type or the other but to determine the degree of each type. Since a blog network tends to exhibit both propensities to a certain degree, it would be more valuable for business to know the degree of propensities of a network.

Girvan and Newman addressed the problem of detecting clusters from a social network [10]. Clusters are sub-networks within which node connections are dense, but between which such connections are much less dense. They proposed a method that detects clusters by exploiting the fact that there are weak connections between clusters [12]. A cluster obtained by this method is a

blog sub-network that is likely to have strong propensity of a particular type. Their method uses only the topology of a network. The strength of the propensity of a blog network is determined by the behavioral pattern (i.e., activities) rather than the structural one, and thus, this method does not provide a good solution to our problem.

Cai *et al.* addressed the problem of mining hidden communities in a social network with multiple heterogeneous relations [6]. They proposed a method that evaluates the importance of different relations contributing to forming a given community. They assumed that the members in the social network have various types of relations and that each relation shows a degree. Their concept may correspond to the two different propensities and the propensity values in our research. Thus, our method for measuring the propensity values can provide their research with various types of relations and their degree it requires.

Lim *et al.* proposed a method for finding bloggers who exhibits high influence in a blogosphere [18]. They measured the influence of a blogger through the number of trackback and scrap actions. The influence of a blogger is similar to IoP, since IoP is computed using the activity patterns of bloggers for diffusing the information. The authors, however, did not distinguish the purpose of use of actions. Furthermore, we measure the degree of the propensities of a network (and not bloggers).

The research problems mentioned above are relevant to but different from the problem we address in this paper. In the following section, we propose a new solution to the problem of determining the degree of the propensities in a blog network.

4. Our Approach

4.1. Overview

Bloggers have both the IoP and the Fop and exhibit either the IoP or the FoP, depending on their counterpart. Thus, we should compute the two propensity values of a relation between two blogs, and using them to determine the propensity values of the whole network.

The propensity values of a relation can be best judged by the blogger who is involved in that relation. We cannot (and should not), however, ask everyone in the network to rate his/her relations every time when the propensities of the blog network need to be analyzed. In order to automate the propensity judgment, we have used a classification technique based on decision trees [14]. For training and testing, we collect a data set through survey and use it to construct two classification models, one for IoP and the other for FoP, respectively. The resulting classification models are used to automatic the process of computing determine the propensity values.

4.2. Determining the Propensity Values of a Relation

4.2.1. Survey. To determine the propensity values of a relation, we conducted a survey. The survey is designed to ask the blogger to rate the propensities of his/her relation. Since the blogger's evaluation can be subjective, the survey includes two sets of sample posts that are carefully chosen to exhibit strong FoP and IoP, respectively. Each survey question asks the blogger how frequently he/she has seen the post similar to the sample post from the blog connected through the relation under question. Our survey contains six questions, three for FoP

and three for IoP. The survey results are quantified and used for computing the propensity values of the relation. The questionnaire used in the survey (translated from Korean) is shown in Appendix.

4.2.2. Class Labels. The answer to each survey question is assigned to a score to capture the degree of a particular propensity. Usually, scores are assigned by domain experts and may be different based on applications. For example, the scores used in our experiments are given in Table 1. As shown in Table 2, the IoP and FoP scores of a relation are computed by adding up the scores of the answers to all the IoP and FoP related questions, respectively.

<Insert Table 1 here.>

<Insert Table 2 here.>

The scores are then transformed into class labels. In our classification model, class labels represent the degrees of the propensity of a relation. We use different numbers of class labels (3, 5, 7, and 9) in our experiments. Depending on applications, more fine-grained class labels can also be used.

4.2.3. Attributes. Among many classification methods, we use classification by decision trees since it is easy to interpret the meaning of classification results.

When building a decision tree, careful selection of attributes is a must [14]. This means, to classify the propensity values of a relation, we need a set of attributes that best represent the characteristics and properties of a relation. Unfortunately, the blog network keeps data about blogs (e.g., total number of posts in a blog, blogger information) and about posts (e.g., total

number of comments on a post, creation time of a post) but not about relations. The attributes of a relation, therefore, needed to be generated out of the data about two blogs connected through the relation under question.

We used one of the most popular blog networks in Korea for building the classification model and running experiments. In this particular blog network, we have identified a set of 30 attributes that seemed to best represent the propensity of a relation, such as the total number of comments made by the parties involved in the relation under question, the number of 'gifts' exchanged between the two, and the number of comments left on the guestbook, etc.

4.2.4. Classification Models. Using class labels and attributes identified in the previous sections, we build two decision trees: one for IoP and the other for FoP. Figure 3 depicts a part of the decision tree for determining the class label for the FoP of a relation. On top is the total number of visits to the other blog, the most important attribute when determining the class label for FoP. The second important attribute is the total number of neighbors. (The list of neighbors is similar to 'blogroll.') We have found that a friendship-oriented blogger tends to keep a small number of acquaintances, whereas an information-oriented blogger tends to have a larger number of neighbors to make it easy to keep track of good information sources. Given that the blog network does not keep explicit data about the relation, the number of neighbors of bloggers involved in the relation under question serves as an important attribute for FoP.

<Insert Figure 3 here.>

Figure 4 depicts a part of the decision tree for determining the IoP of a relation. As in the case with Figure 3, on top is the total number of visits. The second important one, on the other hand,

is the total number of comments the blogger writes. When the blogger is interested in collecting information-oriented contents and sharing them, he/she tends to make a link to the post of his/her interest and put comments on it. Therefore, the number of comments is an important attribute in determining the class label for the IoP.

<Insert Figure 4 here.>

4.3. Determining the Propensity Values of a Network

To compute the degree of the propensities of a blog network, we assign representative values to class labels of the relation, and by adding up the values of all the relations in a network, we compute the propensity values of the whole network.

Simple summation, however, results in the propensity value of the network that depends on its size. Figure 5 exemplifies the potential problem. In Figure 5, nodes represent blogs and edges represent relations. Suppose the propensity value of each relation is 10. Simple summation would result in the propensity value of network B higher than that of network A, since B has more edges than A. To solve the problem that the propensity value of a network tends to increase with the increase in network size, we propose two normalization methods: one normalized over the number of blogs in the network and the other normalized over the number of edges in the network.

<Insert Figure 5 here.>

In the first normalization method, the sum of the propensity values of relations is divided by the total number of blogs in the network. This computes the propensity value per blog on average. Of course, the normalized propensity value tends to be higher when the number of nodes (blogs)

is lower and the number of edges (relations) is higher. In Figure 5, since blog network *A* has 4 nodes and the sum of the propensity values of relations is 60, the propensity value of the network normalized over the number of nodes is 15. On the other hand, blog network *B* has 10 nodes and the sum of the propensity values is 100, which results in the normalized propensity strength of 10.

In the second normalization method, the sum of propensity values is divided by the total number of relations in the network. This computes the propensity value per relation on average. The propensity value normalized over the number of relations tends to be higher when there are edges (relations) with high propensity values. In Figure 5, network *A* has 6 edges and the sum of propensity values is 60, so the propensity value normalized over the number of relations is 10. On the other hand, network *B* has 10 edges and the sum of propensity values is 100, which results in the normalized propensity value of 10.

Table 3 summarizes two normalization methods for determining the propensity values of a blog network. Which normalization methods to use depends on the characteristics of business applications.

<Insert Table 3 here.>

5. Performance Evaluation

In this section, we demonstrate the validity of our approach through extensive experiments using real-world blog data. Section 5.1 describes the experimental setup, and Section 5.2 analyzes the performance of our method.

5.1. Experimental Setup

For experiments, we used anonymized data collected from one of the largest blogospheres in Korea for seven months starting from April 2006. The total number of blogs was about 1,000,000 and the total number of relations (defined as the connection from a blog to a blog since the blog-service provider was mainly interested in the relationships among blogs) was about 2,000,000. Among them, we randomly selected 35,000 relations that showed some level of activities during the experimental period and conducted an online survey on the bloggers involved in those 35,000 relations. To remove invalid data, we compared the answers to the survey and the blog activities recorded. If the survey did not corroborate the recorded activities, we discarded the invalid survey result. If the survey reported frequent visits while the records showed few visits, for example, the relation was discarded from the pool of a data set. This resulted in 1,466 valid relations among 2,801 blogs for analysis.

The proposed method was verified in two ways: the accuracy of the propensity values of a relation and the validity of the propensity values of a blog network.

To measure the accuracy of classification for the propensities of a relation, we used two metric. First, we measured the hit ratio, the percent of samples in the test data set that are correctly classified by the model. Second, we measured the difference, the spread between the class label predicted by the model and the actual class label. Measuring the difference between the correct answer and the predicted one is useful since classifying "High" relation as "Medium" is not as bad as classifying it as "Low."

To measure the hit ratio and the difference, we divided the data set into ten disjoint sets. Nine of them were used as a training data set and the remaining set was used as a testing data set. The selection of nine sets was done in turn, so we were able to perform total of 10 sets of experiments, and the average of the results were used.

Since the actual propensity values of a whole blog network are unknown, showing the performance of our method in determining the propensity values of a network was more challenging. We compared the propensity values of two sub-networks: one that seemed to display a particular propensity more strongly and a randomly-selected network. By showing the network with a strong propensity showed a higher propensity value than that of the randomly-selected one, we were able to suggest the validity of our approach.

To extract a sub-network, we did the following. The network with strong IoP tends to display star topology since bloggers who value information tend to form direct connections to the blogs which maintain comprehensive or up-to-date information. Also, information-seeking bloggers tend not to have connections with blogs that have the same posts as theirs, which results in fewer relations [13]. Therefore, to extract a network that seems to have strong IoP, we start by selecting a hub, the blog with the highest number of trackback links to its posts, and continue to select the blogs that are connected to the hub through trackback. When expanding the network, the blog with a higher number of visits to the hub is included first.

The network with strong FoP tends to have mutual connections. Bloggers who value friendship are likely to have mutual relations with their friends, and a friend of my friend is likely to be a friend of mine, both of which result in a network dense with many relations [22]. To extract a network that is likely to have strong FoP, therefore, we start from the blog with the

highest number of mutual relations with other blogs as a hub. Then, we expand the network by selecting, among the blogs who have mutual relations with the hub, the one with frequent interactions. When extracting a random network, we start from a randomly-selected blog as a hub and include a blog to the network as long as it has a relation with the hub. Using the above method, we extracted 30 networks with strong IoP, 30 networks with strong FoP, and 30 random networks, each of which had 100 blogs in it.

5.2. Analysis

The quality of classification is evaluated by examining whether a known class label of a sample in the testing data set is identical to that predicted from the classification model. In our case, we evaluate whether the class label of the propensity of a relation determined by our decision trees is the same as that obtained by the survey.

5.2.1. Accuracy of the Propensity Values of a Relation.

5.2.1.1 Hit Ratio. The hit ratio measures the percent of samples in the test data set that are correctly classified by the model. Table 4 summarizes the experimental results. The number of labels represents the number of class labels where the propensity scores are grouped into. For example, the 3 labels indicate the propensity scores are grouped into {High, Medium, Low}. As shown in the table, the classification for both IoP and FoP shows the highest hit ratio of 93% when the number of labels is 3, but the hit ratio decreases with the increase in the number of

class labels. Since the score distributions were the same, it is expected that the hit ratio decreases as the number of labels increases.

<Insert Table 4 here.>

5.2.1.2 Difference between Predicted and Actual Propensity Values. Table 5 shows the difference between the predicted class and the actual one. Depending on the number of labels, the difference between the predicted class and the actual class for FoP is between 0.06 and 0.51, and the difference for IoP is between 0.07 and 0.61, respectively. The difference of 0.61, for example indicates the classification is off of 0.61 label on average. Although the difference increases with the increase in the number of labels, the increase is rather small. This result with the hit ratio reported in Table 4, combined verifies the validity of our method, since the difference is rather small even when the hit ratio decreases with the increasing number of labels.

<Insert Table 5 here.>

5.2.2. Validity of the Propensity Values of a Blog Network. Table 6 compares the FoP value of a network with strong FoP and that of a randomly-extracted network. Regardless of the normalization methods used, the network with a strong FoP is determined to have a higher FoP value than the random network.

<Insert Table 6 here.>

Although the FoP value of the network with strong FoP seems higher in both normalization methods, we were not able to determine whether the difference is statistically significant. We

performed T-test for further analysis [15]. The results of the T-test revealed that the difference based on the normalization method over the number of blogs was statistically significant at the 0.05 level of significance. On the other hand, the difference based on the normalization method over the number of relations was not statistically significant. This can be explained by the fact that the bloggers in the network with strong FoP tend to have small number relations and the FoP value of an individual relation is not that high. To determine the FoP value of a network, therefore, we recommend using the normalization method over the number of blogs.

Table 7 compares the IoP values of a network with a strong IoP and of a randomly-extracted network. Regardless of the normalization methods used, the network with a strong IoP is determined to have a higher IoP value than the random network.

<Insert Table 7 here.>

Again, we performed T-test to check whether the difference is statistically significant. The difference based on the normalization method over the number of blogs was not statistically significant at the 0.05 level of significance. The difference based on the normalization method over the number of relations was statistically significant. The blogger whose primary interest is in information collection and dissemination does not keep relations with blogs that have the same information. He/She probably wants to connect to a few blogs which keep the most up-to-date or most comprehensive information. Therefore, the network with strong IoP is likely to be organized into a collection of star topology, and the network itself is likely to have fewer relations than the random network. The IoP value of such a network, if computed with the normalization over the number of blogs, tends to be lower. When determining the IoP value of a blog network, therefore, we recommend the normalization method over the number of relations.

6. Conclusion

A blog network exhibits two propensities based on the purpose of use: Information-oriented Propensity and Friendship-oriented Propensity. Knowing the relative and absolute values of these propensities of a blog network serves as an important basis for business and policy decisions of how to utilize the network. This paper proposes a new approach to judge the propensity values of a blog network.

We divide the problem into two pieces: determination of the propensity values of a relation and those of a network. We collected a data set through survey for training and testing in classification, and using them, we have developed an automated method to determine the propensity values of a relation. The propensity values of a network are determined by adding up the representative values of all relations in the network. We have also proposed two normalization methods to prevent the problem that the propensity value tends to depend on the size of the network.

The hit ratio of the proposed method decreases from 93% to 60% and the difference increases from 0.06 to 0.61 with the increase in the number of class labels. The increase in difference is relatively small compared to the decrease in the hit ratio, which validates the effectiveness of the proposed method of determining the propensity values of a relation.

Since no direct measurement of the propensity values of a network is available, we devise a heuristic for extracting sub-networks seemingly having strong propensities, and compare their propensity values with those of randomly-selected networks. The analysis strongly suggests that

our approach is well suited for measuring the propensity values of a network. The statistical analysis done through T-test also reveals that the normalization method over the number of blogs is appropriate for measuring FoP, and the normalization method over the number of relations is appropriate for measuring IoP.

The contribution of this paper is that it provides an automated and quantitative method that can be used to judge the propensity values of a blog network, which would be useful in business and policy decisions for blog-related applications. Although the paper focuses on two propensities, IoP and FoP, our method can be extended and applied to judging other types of properties of a blog network.

References

1. Agrawal, R.; Rajagopalan, S.; Srikant, R.; and Xu, Y. Mining Newsgroups using Networks Arising from Social Behavior. In *Proc. Int'l. Conf. on World Wide Web*, 2003, pp. 529-535.
2. Balasubramanian, S. and Mahajan, V. The Economic Leverage of the Virtual Community. *International Journal of Electronic Commerce*, 5, 3 (Spring 2001), 103–138.
3. Barabasi, A. *Linked: the new science of networks.* New York: Perseus Books Group, 2002.
4. Boyd, D. and Ellison, N. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*, 13, 1 (January 2007), from http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html.
5. Brown, J. and Reinegen, P. Social Ties and Word-of-Mouth Referral Behavior. *Journal of Consumer Research*, 14, 3 (December 1987), 350-362.

6. Cai, D.; Shao, Z.; He, H.; Yan, X.; and Han, J. Mining Hidden Community in Heterogeneous Social Networks. In *Proc. Int'l Workshop on Link Discovery*, 2005, pp. 58-65.
7. Chin, A. and Chignell, M. A Social Hypertext Model for Finding Community in Blogs. In *Proc. Int'l. Conf. on Hypertext and Hypermedia*, 2006, pp. 11-22.
8. Domingos, P. and Richardson, M. Mining the Network Value of Customers. In *Proc. Int'l. Conf. on Knowledge Discovery and Data Mining*, 2001, pp. 57-66.
9. Duncan W. and Peter D. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer research*, 34, 4 (2007), 441-458.
10. Girvan, M. and Newman, M. Community Structure in Social and Biological Networks. *Proceedings of the National Academic of Science*, 99 (June 2002), 7821-7826.
11. Goldenberg, J., Libai, B. and Muller, E. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, 12, 3 (August 2001), 211-223.
12. Granovetter, M. The strength of weak ties. *American Journal of Sociology*, 78, 6 (May 1973), 1360-1380.
13. Gruhl, D.; Guha, R.; Liben-Nowell, D.; and Tomkins, A. Information Diffusion through Blogspace. In *Proc. Int'l. Conf. on World Wide Web*, 2004, pp. 491-501.
14. Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*. California: Academic Press, 2001.
15. Hogg, R.; and Tanis, E. *Probability and Statistical Inference*. New Jersey: Prentice Hall, 1996.
16. Koh, J. and Kim, Y. Sense of Virtual Community: a Conceptual Framework and Empirical Validation. *International Journal of Electronic Commerce*, 8, 2 (Winter 2003), 75–93.

17. Lechner, U. and Hummel, J. Business Models and System Architectures of Virtual Communities: from a Sociological Phenomenon to Peer-to-peer Architectures. *International Journal of Electronic Commerce*, 6, 3 (Spring 2002), 41–53.
18. Lim, S.; Kim S.; Park S.; and Lee J. Determining Content Power Users in a Blog Network. In *Proc. Int'l Workshop on Social Network Mining and Analysis*, 2009.
19. Minjae C. *Blog Industries in Korea*. Seoul: Korea Press Foundation, 2009 (in Korean).
20. Scoble, R.; and Israel, S. *Naked Conversations*, New York: John Wiley & Sons Inc, 2006.
21. Wasserman, S. and Faust, K. *Social Network Analysis: Methods and Applications*, New York: Cambridge University Press, 1994.
22. WegoNet. *Brand Strategy in Communities*. Seoul: E-Design Press, 2004 (in Korean).
23. Wright, J. *Blog Marketing*. New Jersey: McGraw-Hill, 2005.

Appendix: The Survey Questionnaire

(A sample post that exhibits strong FoP is shown.)	
1	How often have you seen posts related to daily life (e.g. diary, essays, personal stories, etc.) from blog A?
2	How often have you seen 'personal' photographs (e.g. his/hers, his/her friends', his/her family's, etc.) from blog A?
3	How often have you left comments, replies and posts of greetings (e.g. posts in guestbook, comments to say hello, etc.) on blog A?
(A sample post that exhibits strong IoP is shown.)	
1	How often have you seen posts providing useful information (e.g. book reviews, movie reviews, stock information, restaurant guide, product information, etc.) from blog A?
2	How often have you seen photographs related to specific topics (e.g. movie posters, sports, arts, animations, etc.) from blog A?
3	How often have you seen funny, useful or interesting posts (e.g. personality test, articles about special topics, etc.) from blog A?

Figures

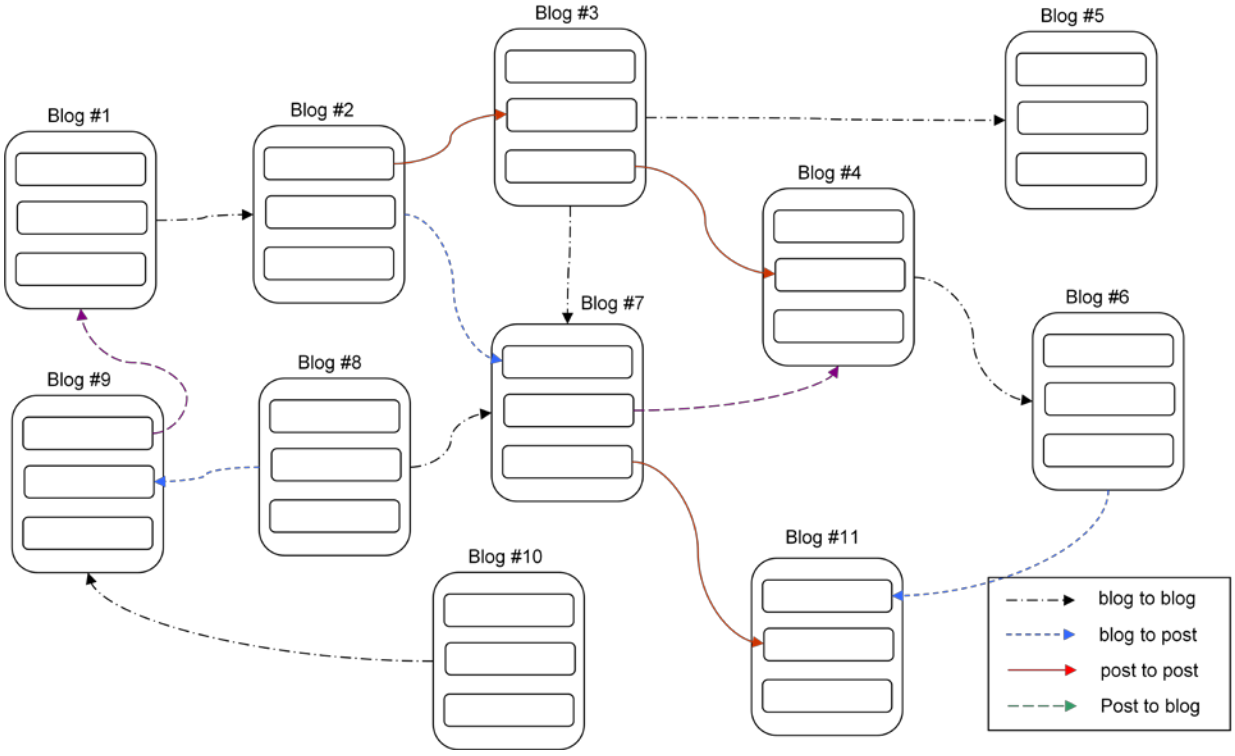


Figure 1. "An example of a blogosphere"

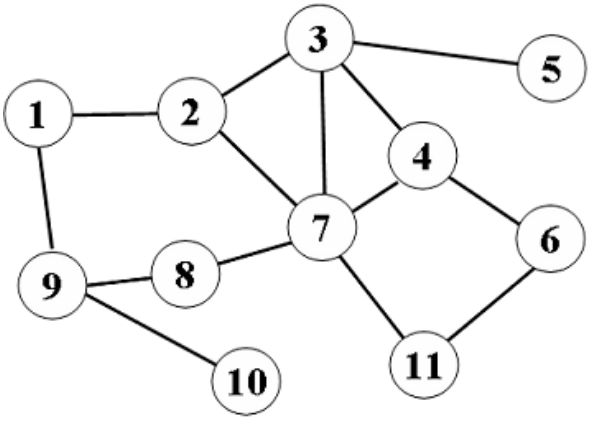


Figure 2. "The network representation of the blogosphere in Figure 1"

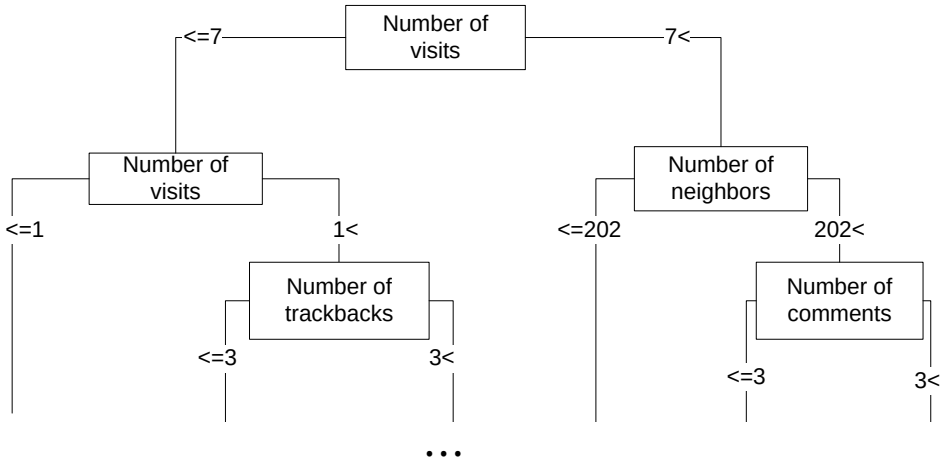


Figure 3. “Decision tree for the FoP of a relation”

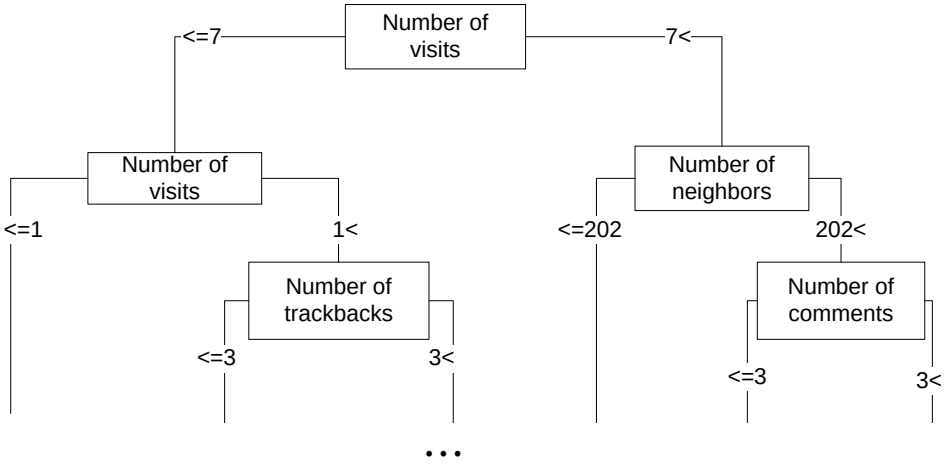
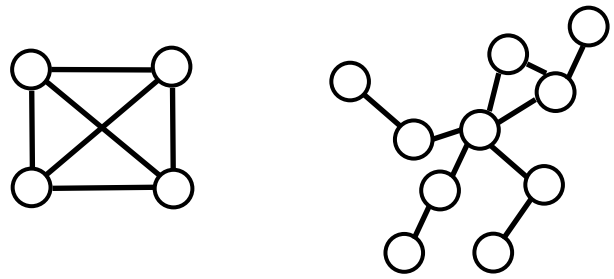


Figure 4. “Decision tree for the IoP of a relation”



Blog network A
(Propensity value is 60)

Blog network B
(Propensity value is 100)

Figure 5. "The propensity value of the network based on simple summation"

Tables

Table 1. "The scores assigned to the answers to the question of the frequency of reading a post similar to the sample post"

" More than once a day "	: 30
" More than once a week "	: 15
" More than once a month "	: 5
" More than once or twice total "	: 1
" Never "	: 0

Table 2. "Computation of propensity scores from survey answers"

IoP	$PE_I = \sum_{i}^{N(Q_I)} S_i$
FoP	$PE_F = \sum_{i}^{N(Q_F)} S_i$
Notations	i : Question number S_i : Score of question i PE_I : Score of IoP PE_F : Score of FoP $N(Q_I)$: Number of questions related to IoP $N(Q_F)$: Number of questions related to FoP

Table 3. "Two normalization methods"

Propensity Type / Normalization Basis	IoP	FoP
Number of Blogs	$PN_I = \frac{\sum PE_I}{N_v}$	$PN_F = \frac{\sum PE_F}{N_v}$
Number of Relations	$PN_I = \frac{\sum PE_I}{N_e}$	$PN_F = \frac{\sum PE_F}{N_e}$
Notations	N_v : Number of Nodes(Blogs) N_e : Number of Edges(Relations) PN_F : FoP of a Network PN_I : IoP of a Network PE_F : FoP of a Relation PE_I : IoP of a Relation	

Tabel 4. "Hit ratio"

Types / Number of Labels	FoP	IoP
3	94%	93%
5	74%	70%
7	63%	61%
9	60%	66%

Table 5. "Difference between the predicted class label and the actual class label"

Types / Number of Labels	FoP	IoP
3	0.06	0.07
5	0.27	0.35
7	0.43	0.55
9	0.51	0.61

Table 6. "Comparisons of FoP values of a network"

Network types / Normalization basis	Network with Strong FoP	Random Network
Number of Blogs	3.26	2.18
Number of Relations	1.17	1.15

Tabel 7. "Comparisons of IoP values of a network"

Network types / Normalization basis	Network with Strong IoP	Random Network
Number of Blogs	2.51	2.18
Number of Relations	2.27	1.15