Regular Paper

Quantifying Influence in Social Networks and News Media

Hongwon Yun, *Member*, *KIICE*

Department   of Information Technology, Silla University, Busan 617-736, Korea

Abstract

Massive numbers of users of social networks share various types of information such as opinions, news, and ideas in real time. As a new form of social network, Twitter is a particularly useful information source. Studying influence can help us better understand the role of social networks. The popularity of social networks like Twitter is primarily measured by the number of followers. The number of followers in Twitter and the number of users exposed to news media are important factors in measuring influence. We chose Twitter and the New York Times as representative media to analyze the influence and present an empirical analysis of these datasets. When the correlation between the number of followers in Twitter and the number of users exposed to the New York Times is computed, the result is moderately high. The correlation between the number of users exposed to the New York Times and the number of sections including the users on it, was found to be very high. We measure the normalized influence score using our proposed expression based on the two correlation coefficients.

Index Terms: Social network, Twitter, Influence, Influence score

I. INTRODUCTION

Currently, one of the most notable micro-blogging services is Twitter. Twitter is used by many people as a tool for spreading their ideas, knowledge, or opinions to others. Users in Twitter are usually dubbed "twitterers," and they can publish "tweets" through a variety of media to others. Twitter has been gaining huge popularity and also increased interest from researchers [1-4]. In recent years, interest among researchers has increasingly focused on whether or not influential people have the power to influence a large number of others in social networks [5-10]. Recently, Twitter users and applications have been considering a twitterer's influence to be measured by the number of followers the twitterer has [11-13]. The popularity of a twitterer depends on the number of followers [6, 9]. The question is, are people with many followers on Twitter influential people in their community? How many times have they have they received coverage in the news media? We are interested in identifying people who are influential in both social networks and the news media.

In this paper, we quantify the influence on Twitter and the news media based on influential twitterers and present an empirical analysis. To do this work, we gathered top users' data from Twitter based on the number of followers. We then searched news articles in the online news media using keywords that were collected from top users' names from Twitter. These two datasets gathered from Twitter and the news media were analyzed to find a good indicator of influence. To investigate the correlation between influential twitterers who were also influential in the news media, we evaluated the correlation coefficient. Using these values, the influence scores were measured by our proposed approach in order to obtain the value of influence on Twitter and in the news media.

The rest of this paper is organized as follows. The

_____

datasets were prepared for the purpose of this study. We provide an overview of the collected data and show the preliminaries in section II. Section III elaborates on the analysis of data and the methodology for estimating user influence. In section IV, we present the quantifying influence score and the empirical results. Finally, we summarize our research and suggest some directions for future work in section V.

II. DATA PREPARATION AND PRELIMINARIES

For the purpose of this study, a set of Twitter data was prepared on January 27, 2012. We collected the global top 100 users on Twitter. The global top 100 users have the largest number of followers worldwide. For the Twitter collections, we have removed all the organizations to select only individuals as influential twitterers. The top twitterers were ranked by the number of followers. After removing all organizations from top 100 users in Twitter, 80 individuals remained. We use the 80 individuals as our basic dataset in this study. In order to obtain the total number of news articles published that included each name in the global top 80 twitterers, we chose the New York Times (NYT) as our sample. We gathered news articles dating from January 28, 2011 to January 27, 2012 through the NYT's search page.
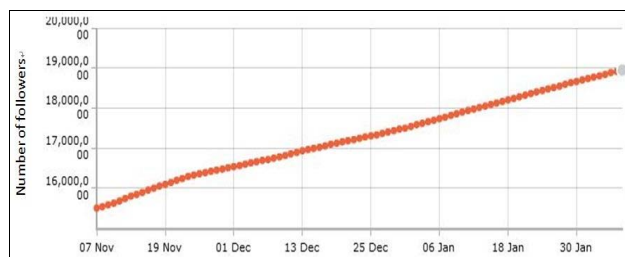
(a) Lady Gaga
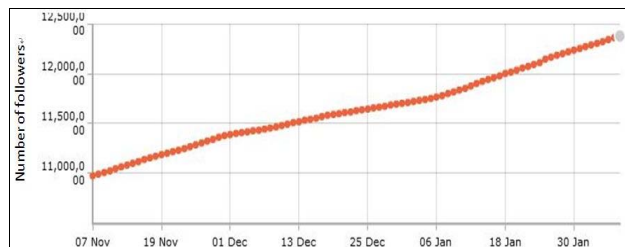
(b) Barack Obama

Fig. 1. Rapid growth of the top twitterers' followers.

As an example, it may be worthwhile to consider how many followers Lady Gaga lost or gained. With regard to the number of followers, the statistics can help us to

discover a twitterer's influence. Fig. 1 shows the rapidly increasing number of the most popular twitterer's followers. For example, the number of Lady Gaga's followers was 15,685,045 on November 11, 2011 and after 3 months it was 19,030,903 on February 10, 2012. Over those 3 months, the number of her followers increased by 3,345,858 as shown in Fig. 1a. Fig. 1b shows that a similar increase appeared in the case of President Barack Obama of the United States.

The top 5 users on Twitter ranked by the number of followers are shown in Table 1. The number following and tweets are also shown in Table 1. Even though all of the top 5 users have a large number of followers, they follow few others and have few tweets. There is a large difference between the number of followers and the number following and number of tweets. Fig. 2 shows the top Twitter users who are ranked by the cumulative number of followers, following, and tweets. As shown in this figure, the number of followers captures almost entirely the cumulative distribution.

Table 1. Units for magnetic properties top 5 twitters ranked by the number of followers

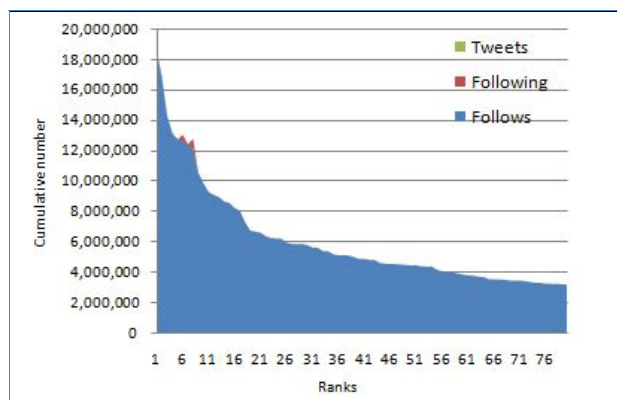| Rank	Name	No. of followers	No. following	No. of tweets
1	Lady Gaga	18,514,135	140,108	1,231
2	Justin Bieber	16,788,783	120,698	12,946
3	Katy Perry	14,325,345	82	3,668
4	Shakira	13,174,664	43	1,234
5	Kim Kardashian	12,793,334	167	10,175

Fig. 2. Top twitterers ranked by the cumulative number of followers, following, and tweets in Twitter.

III. DATASET ANALYSIS

To understand the features of the datasets from Twitter and the NYT, we first analyze the datasets.

A. Basic Analysis of Twitter Data

Fig. 3 displays the number following against the number of followers in Twitter; the inside figure shows the number following around 5,000,000 followers in detail. Barack Obama is ranked 8 by the number of followers, and has the most number of 683,249 following. On the other hand, Marshall Bruce Mathers III (Eminem) has the least number others he is following, and he is ranked 15 by the number of followers.
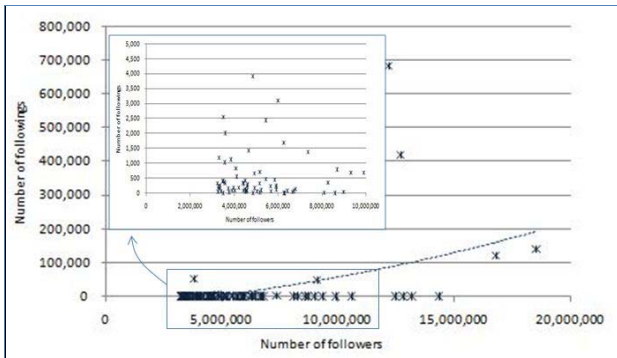

Fig. 3. Number of followers and number following for the top twitterers in Twitter.
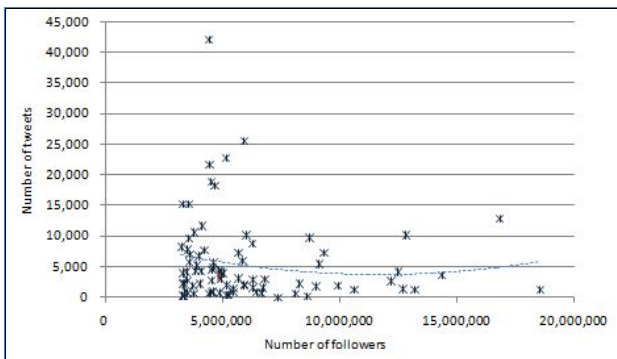

Fig. 4. Number of followers and number of tweets for the top twitterers.
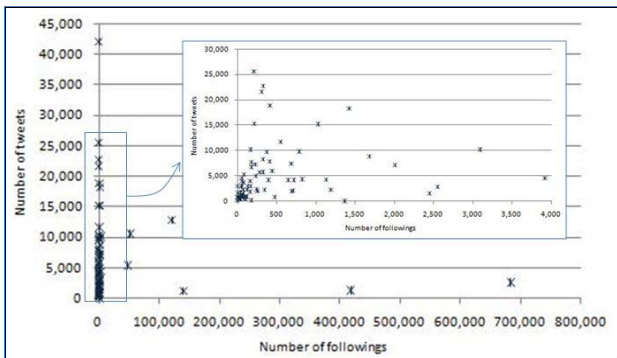

Fig. 5. Number following and number of tweets for the top twitterers

We plot the distribution between the number of followers and that of tweets in Fig. 4. Fig. 5 shows the number of tweets against the number following; the inside magnified figure displays a scale of around 500 following. We can see that there is little correlation between the number of followers, the number of others following, and the number of tweets as shown in Figs. 3-5 and as pointed out previously.

B. Basic Analysis of NYT Data

For each of the top 80 twitterers, in order to investigate the number of news articles covering them in the NYT, we queried by the name of each of the top 80 twitterers. Table 2 shows the number of news articles for each section. As we can see from the table, the largest number of news articles that covers the 80 twitterers is found in the Arts section. This means that many popular Twitter users are engaged in the arts. We can see a major difference between the total number of news article in the Arts section and that of others as shown in Fig. 6.

Table 2. Proportion of news articles in the New York Times queried by the top 80 twitters

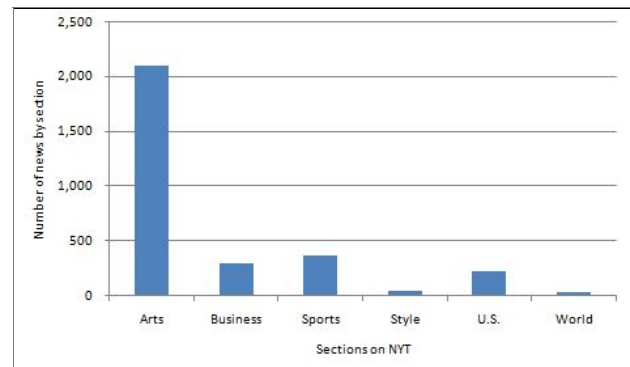|	Arts	Business	Sports	Style	U.S.	World
No. of articles	2,100	299	374	49	231	41
% of articles	68	10	12	2	7	1

Fig. 6. Distribution of the number of news articles in each section in the New York Times (NYT) covering the top twitterers.

C. Methodology for Estimating User Influence

1) The Spearman's Rank Correlation Coefficient

It is not easy to determine whether influential twitterers are also influential or not, even in the news media. Correlations are useful because they can indicate a

predictive relationship that can be exploited in practice. In order to quantify correlation, we computed the correlation coefficient between the two different datasets. We used the relative order of users' ranks as a value of the difference. Each value was sorted, so that the rank of 1 means the most influential user and increasing rank indicates a less influential user. Every user is assigned a rank for each influence measure. Here the Spearman's rank correlation coefficient can quantify a user's rank varies across different values. We use the Spearman's rank correlation coefficient as a measure of the strength of the association between two rank sets:

$$\rho = 1 - \frac{6\sum(x_i - y_i)^2}{N(N^2 - 1)}$$

(1)

The variables x_i and y_i are the ranks of users based on two different influence measures in a dataset n. The closer ρ is to +1 or -1, the stronger the likely correlation. A perfect positive correlation is +1 and a perfect negative correlation is -1 [14, 15].

2) Influence Score

We can compute the influence scores of the users based on the Spearman's rank correlation coefficient. To do so, we define the influence score (S) as follows:

$$S = \left(1 - \frac{\sqrt{r_i} \times \sqrt{r_j}}{N} \times \rho_{(i,j)}\right) \times \left(1 - \frac{\sqrt{r_k} \times \sqrt{r_l}}{N} \times \rho_{(k,l)}\right)$$

(2)

This value S represents the amount of influence between two pairs of ranked datasets. We call this value the influence score for the same domain. A value r_i is a sorted rank within a dataset. The value r_j represents is also ranked one within a same dataset. The correlation coefficient $\rho_{(i,j)}$ is a measure of the strength of the association between two ranked datasets. The variable N is a sample size. Likewise, r_k and r_l are the ranked value within another dataset. The correlation coefficient $\rho_{(k,l)}$ represents a measure of statistical dependence between another two ranked datasets. We use this equation to evaluate the correlation in section IV.

IV. QUANTIFYING INFLUENCE SCORES

We describe the correlation between the top users' ranking in Twitter and their ranking in the NYT. Normalized influence scores are computed for each of the top twitterers and each of them in the NYT.

A. Distribution of the Top Twitter Users and Exposure in NYT Articles

The top 80 twitterers ranked by the number of followers are showed in Fig. 7a. We found the number of news articles that mentioned the names of the top 80 users and sorted the number of new articles. Many of the people who are exposed in the NYT relatively frequently can be referred to as celebrities. The distribution of the celebrities is shown in Fig. 7b.
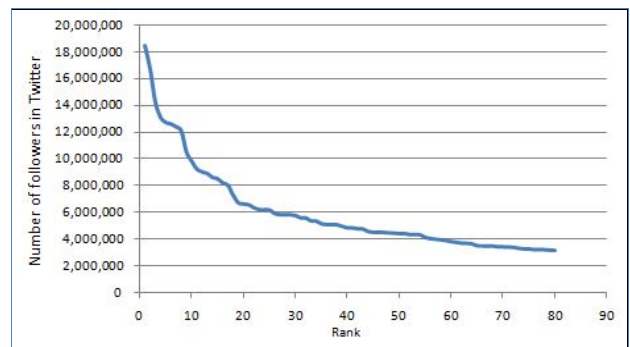
(a) Distribution of number of followers for the top twitterers
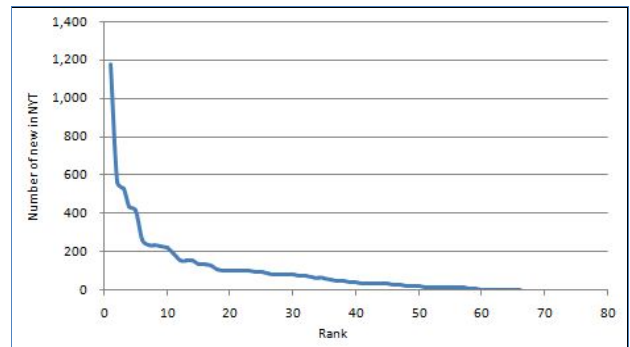
(b) Exposure degrees of news articles in New York Times based on the top twitterers

Fig. 7. Number of top users in Twitter and exposure of news articles in the New York Times (NYT).

B. Comparing Twitter and NYT based on the Top Users in Twitter

The top 5 users by the number of followers and corresponding rankings by the number of news articles are listed in Table 3. We can see that the high rankings in Twitter are not necessarily high in the NYT in Table 3. Fig. 8 shows the number of news articles against the ranking by the number of followers. This figure means that the popularity of users in Twitter is not proportionate to the amount of exposure in the NYT.

Table. 3. Comparison of the top users between Twitter and New York Times

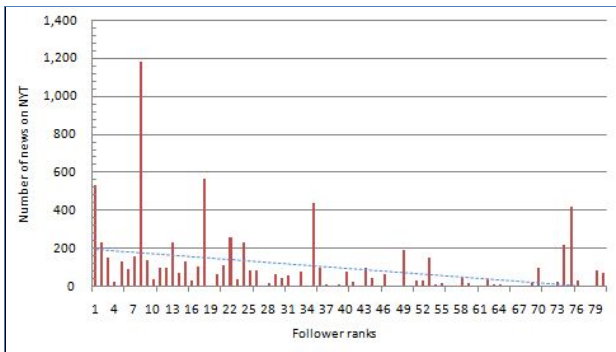| Twitter		New York Times		
Rank	No. of followers	Rank	No. of news	No. of sections
1	18,514,135	3	534	24
2	16,788,783	7	236	22
3	14,325,345	13	155	18
4	13,174,664	47	29	9
5	12,793,334	17	131	20

Fig. 8. Followers' rankings for the top users in Twitter and number of news articles searched by each of them in the New York Times (NYT).

C. Estimating Correlation Coefficient

In order to investigate how the two pairs correlate, we compare the relative influence ranks of all 80 users. We compute the correlation coefficient as shown in Table 4 by the Spearman's rank correlation coefficient. The value 0.52 in Table 4 is the correlation coefficient between the number of followers in Twitter and the number of news articles in the NYT. We see a moderately high correlation (above 0.5) across all pairs.

Table. 4. Spearmans's rank correlation coefficient

Correlation	Correlation coefficient
No. of followers in Twitter vs. No. of articles in New York Times	0.52
No. of articles in New York Times vs. No. of sections in New York Times	0.93

D. Measuring Influence Score on Twitter and NYT

The popularity of a Twitter user can be easily measured by the number of followers. However, the number of followers in Twitter alone cannot be a measure for estimating influence. We computed the influence score using our proposed expression as defined in section III. The

top 20 users by the influence score are listed in Table 5. Lady Gaga ranks 1 after measuring the influence score and Barack Obama's ranking went up in the second place due to the large amount of exposure in the NYT.

Table. 5. Top 20 users ranked by # of followers, top 20 users ranked by number of articles and top 20 celebrities ranked by influence score

Rank	Twitter		NYT		Twitter & NYT	
	Name	No. of followers	Name	No. of news	Name	Influence score
1	Lady Gaga	18,514,135	Barack Obama	1,183	Lady Gaga	0.45
2	Justin Bieber	16,788,783	Chris Brown	565	Barack Obama	0.45
3	Katy Perry	14,325,345	Lady Gaga	534	Chris Brown	0.43
4	Shakira	13,174,664	Bill Gates	438	Justin Bieber	0.42
5	Kim Kardashian	12,793,334	LeBron James	418	Bill Gates	0.40
6	Britney Spears	12,675,440	Charlie Sheen	261	Oprah Winfrey	0.39
7	Rihanna	12,472,675	Justin Bieber	236	Charlie Sheen	0.37
8	Barack Obama	12,165,236	Oprah Winfrey	235	Katy Perry	0.37
9	Taylor Swift	10,587,678	Kanye West	232	Kim Kardashian	0.35
10	Selena Gomez	9,898,228	Adele	221	Kanye West	0.35
11	Ashton Kutcher	9,314,038	Chelsea Handler	193	Rihanna	0.34
12	Ellen DeGeneres	9,099,227	Rihanna	157	Marshall Mathers	0.33
13	Oprah Winfrey	8,979,402	Katy Perry	155	LeBron James	0.32
14	Nicki Minaj	8,692,074	JennifeLopez	154	Britney Spears	0.31
15	Marshall Mathers	8,589,804	Taylor Swift	137	Taylor Swift	0.30
16	Kaka	8,269,395	Marshall Mathers	134	Ashton Kutcher	0.29
17	Justin Timberlake	8,090,049	Kim Kardashian	131	Justin Timberlake	0.28
18	Chris Brown	7,351,342	Cristiano Ronaldo	111	Ellen DeGenre	0.28
19	Bruno Mars	6,710,772	Justin Timberlake	105	Paris Hilton	0.27
20	Cristiano Ronaldo	6,648,611	Ashton Kutcher	103	Adele	0.27

NYT: New York Times.

IV. CONCLUSIONS

Social media has been growing explosively and provides users the opportunity to share various types information and knowledge in real time. Twitter is one of the most popular social networks on the internet. The popularity of these social networks can be measured by the number of followers. The influence, which is, the individual's potential to lead others to engage in a certain act, can also be determined by many factors. For studying influence, we prepared a set of global top twitterers' data and collected news articles mentioning each of their names from the NYT. The number of followers is a significant factor for measuring influence.

The exposed number of users in the news media is an important factor used to estimate the influence. We chose Twitter and the New York Times as representative media to analyze the influence. To understand the features of these two datasets, we analyzed the datasets and presented the results. We show that the correlation between the number of followers and the number of top twitterers with strong exposure in the news media is moderately high. The correlation between the exposed number of top twitterers and the number of sections including them is very high. Our proposed expression using these two correlation coefficients computed the influence score for each of the top twitterers ranked by the number of followers. This normalized influence score opens the possibility for discovering influential individuals within all of the social networks and news media.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, NC, pp. 591-600, 2010.

[2] W. X. Zhao, J. Jiang, J. Weng, J. He, E. P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, Dublin, Ireland, pp. 338-349, 2011.

[3] B. Suh, L. Hong, P. Pirolli, and E. H. Chi, "Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network," *Proceedings of the 2nd IEEE International Conference on Social Computing*, Minneapolis, MN, pp. 177-184, 2010.

[4] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on twitter," *Proceedings of the 20th International Conference on World Wide Web*, Hyderabad, India, pp. 705-714, 2011.

[5] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, Hong Kong, China, pp. 65-74, 2011.

[6] J. Weng, E. P. Lim, J. Jiang, and Q. He, "TwitterRank: finding topic-sensitive influential twitterers," *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, New York, NY, pp. 261-270, 2010.

[7] H. Kwak, H. Chun, and S. Moon, "Fragile online relationship: a first look at unfollow dynamics in twitter," *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, Vancouver, Canada, pp. 1091-1100, 2011.

[8] M. de Choudhury, N. Diakopoulos, and M. Naaman, "Unfolding the event landscape on twitter: classification and exploration of user categories," *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, Seattle, WA, pp. 241-244, 2012.

[9] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," *Proceedings of the 20th International Conference on World Wide Web*, Hyderabad, India, pp. 113-114, 2011.

[10] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "measuring user influence in twitter: the million follower fallacy," *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington, DC, pp. 10-17, 2010.

[11] E. Bakshy, B. Karrer, and L. A. Adamic, "Social influence and the diffusion of user-created content," *Proceedings of the 10th ACM Conference on Electronic Commerce*, Stanford, CA, pp. 325-334, 2009.

[12] E. Fischer and A. R. Reuber, "Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behavior?" *Journal of Business Venturing*, vol. 26, no. 1, pp. 1-18, 2011.

[13] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*, New York, NY, pp. 241-250, 2010.

[14] J. L. Myers and A. Well, *Research Design and Statistical Analysis*, 2nd ed., Mahwah, NJ: Lawrence Erlbaum Association, 2003.

[15] J. S. Maritz, *Distribution-Free Statistical Methods*, London: Chapman and Hall, 1981.

Hong-won Yun

received his B.S. and the Ph.D. degrees from the Department of Computer Science of Pusan National University, Korea in 1986 and 1998, respectively. He is a professor in the Department of Information Technology, Silla University in Korea. His research interests include databases and social networks.