Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 10(1): 27-32, Mar. 2012 **Regular Paper**

Analysis of Similarity of Twitter Topic Categories among Regions

Hongwon Yun*, ***Member, KIICE***

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

Abstract

Twitter can spread and share all kinds of information such as facts, opinions, and ideas in real time. In this paper, we empirically compare and analyze the topic categories in Twitter with all top 100 users in each of geographic region. We mainly consider the relationships among regions and selected four regions: Global, Seoul, Tokyo, and Beijing. Each of the top 100 users in Twitter is classified into a specific category and then statistical analysis is conducted. Among eight topic categories, the "Arts" category is the largest and the second is "Life". The correlation between global and Seoul groups has the lowest value among the six pairs of relationships between regional groups, and this difference is statistically significant. We find that the Seoul, Tokyo, and Beijing regional Twitter groups, all in East Asia, have high topical similarity. Based on the correlation analysis, Seoul and Tokyo saliently show a sticky trend. The correlation coefficient presents very a strong positive correlation between Seoul and Tokyo. The correlation between the global group and the East Asian groups is relatively lower than that among the East Asian groups.

Index Terms: Twitter, Regional similarity, Topic category, Regional trend

I. INTRODUCTION

Twitter is the most popular micro-blogging and social media service and service in which it is possible to share information in real time. Twitter allows its users to send short messages to others and these messages are referred to as "tweets" and can be sent and retrieved through a variety of media. Twitter users are able to send short texts with mobile devices and "retweet" those messages to their followers. A great deal of information is shared across the globe via its real time network. Twitter can spread and share all kinds of information such as news, marketing, real time events, and even their ideas because the tweets are short and convenient.

Twitter provides access to the thoughts, opinions, activities, and experiences of several hundred millions of twitter users in real time, with the option of sharing the user's location. Fig. 1 shows the average number of tweets per second between September and November, 2010. As we can see from the figure, Twitter users send more than 1,000 tweets per second. Twitter as a new form of social network service has been widely adopted for personal and organizational purposes. This leads to dynamic new political, cultural and lifestyle trends. The popularity of Twitter has made this new social media service an attractive object of study. Many researchers have been investigating how these social media and data are used. This rich source of data is motivating a growing body of scientific study of user motivation and collaboration [1-5]. Some researchers have focused specifically on crisis management and

Received 04 December 2011, **Revised** 18 January 2012, **Accepted** 24 January 2012
***Corresponding Author** E-mail: hwyun@silla.ac.kr

Open Access **http://dx.doi.org/10.6109/jicce.2012.10.1.027** print ISSN:2234-8255 online ISSN:2234-8883

collective problem solving in mass emergency events [6-9]. Recently, content analysis on Twitter has been carried out [10-18], and interest in this area has been increasing.

These studies include investigations of the characteristics of social networks, the role of real time sensors, analysis of tweets, and so on. However, the empirical analysis and topical comparison between countries of similar regions such as Seoul, Tokyo, and Beijing in the East Asian region are still in an early stage. Even though a great deal of information is shared via Twitter's social network, little is known about who uses social networks and what topic categories they cover. Studies about regional similarity on Twitter in particular have not been published yet. In this study, we are interested in analyzing the regional topic variation in the content of the top Twitter, focusing on East Asia, specifically Seoul, Tokyo, and Beijing. We are motivated to investigate the feasibility of identifying regional similarities by statistical analysis.

In this paper, we empirically compare the trends in topics on Twitter with the top 100 users in each of the regions. To do this work, we collected the top 100 users in Twitter based on the number of followers. We thoroughly classified and analyzed them statistically. The rest of this paper is organized as follows. First, we present how we gathered data and what data were collected. Preliminary statistical analysis is also presented in section II. Next, each of the top 100 users in Twitter was categorized into specific categories. We introduce the statistical methods we used to analyze the data in this study. In section IV, we present the empirical comparison results focusing on regional similarities. Finally, we summarize our research and suggest some directions for future work in section V.

II. DATA PREPARATION AND PRELIMINARIES

We collected all of the top 100 Twitter users from each region, based on the largest number of followers. For the purpose of analysis by category, we considered regional relationships and selected four regions: Global, Seoul, Tokyo, and Beijing. We sometimes use the term East Asia for the 3 regions of Seoul, Tokyo, and Beijing. The total of all the top users of the four regions together is 400 people [19]. We analyzethe collected dataset for the purpose of understanding the basic statistics of the dataset in Twitter. These preliminary statistical analyses are shown in Figs.2, 3, and Table 1.

According to our investigation, as of October 20, 2011, Lady Gaga of the Global group had 14,694,619 followers. In Korea, 960,464 Twitter users follow Lee Oisoo. Lady Gaga had 15 times as many followers as Lee Oisoo did. More generally, the Global group has a large difference in number of followers compared to the East Asian groups as shown in Fig. 2.

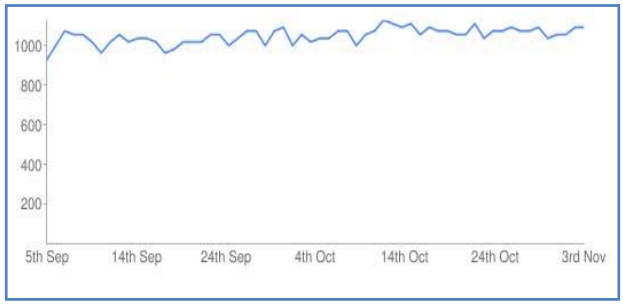

Fig. 1. Average number of tweets per second.

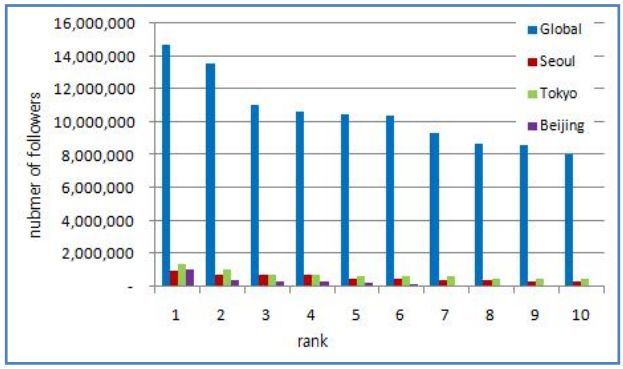

Fig. 2. Number of followers of top 10 Twitter users in Global, Seoul, Tokyo, and Beijing.

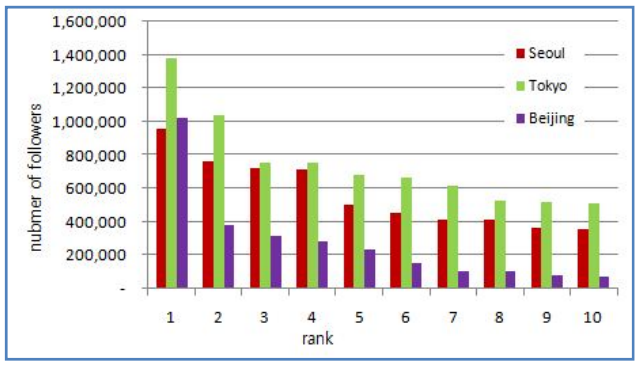

Fig. 3. Number of followers of the top 10 Twitter users in Seoul, Tokyo, and Beijing, with Global excluded.

Table 1. Statistics of top 10 Twitter users

Statistics of followers	Global	Seoul	Tokyo	Beijing
Sum	105,544,518	5,670,706	7,446,123	2,759,283
Standard deviation	2,151,984	207,969	273,202	284,353
Mean	10,554,452	962,448	744,612	275,928
Median	10,440,020	480,397	673,237	194,742
Max	14,694,619	960,464	1,378,069	1,024,691

A comparison for East Asia on a smaller scale is needed. Fig. 3 shows the three major regions in East Asia in more detail. The statistics for the top 10 Twitter users are shown

in Table 1. The number of followers in the US-centric Global group has significant differences from the other regions of Seoul, Tokyo, and Beijing.

III. TOPIC CLASSIFICATION

We categorize topics by assigning a topic to a category and adopt a statistical method to analyze data by topic categories.

A. Categorizing Topics

As shown in Table 2, we categorized topics by assigning a topic to category. Each of the top 100 users of Twitter was classified into a specific category. Basically, we use the classification of browsing interests in Twitter and adjusted some categories based on our own judgment. For example, music and art are classified separately in Twitter. However, in this study, music is classified as part of Arts since it is included in the arts conceived more broadly. First, we examined the topic words each user used in their biography. Next, we focused on the topic words that are classified into a category. In some cases, there were vague and difficult to determine. The results of this work are shown in Table 2.

Table 2. Specific topics in each category

Category	Topics
Arts	Fans, story, artist, writer, singer, author, music, movie, actress, actor, album, novelist, journalist
Business	Facebook, CEO, company, shopping, website, chairman, media, director, foundation, bank, financial
Education	University, professor, teacher, study, researcher
Government	Campaign, president, official, representative, minister, mayor, center, ambassador, governor
Life	Live, posting, happiness, city, camping, life, citizen, health, family, friend, blogger, love, travel, normal, people, laughing, network
News	CNN, NYT, news, Onion, world, breaking, Time, globe, Mainichi, Asahi, NHK, KBS, newspaper
Science&Tech	Science, IT, programmer, researcher, engineer, specialist, coder
Twitter	Spanish, Japanese, Chinese, Twitter, Tweet

B. Statistical Method for Analysis

Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. To analyze the categorized data for all of the top 100 users in each region, we used the Spearman's rank correlation coefficient, which is defined as the Pearson correlation coefficient between the ranked variables. The n raw scores X_i, Y_i converted to ranks x_i, y_i, and ρ are computed from these:

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \quad (1)$$

Tied values are assigned a rank equal to the average of their positions in the ascending order of the values. In applications where ties are known to be absent, a simpler procedure can be used to calculate ρ. Differences $d_i = x_i - y_i$ between the ranks of each observation on the two variables are calculated, and ρ is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2)$$

There are several other numerical measures that quantify the extent of statistical dependence between pairs of observations. It is common to regard these rank correlation coefficients as alternatives to Pearson's coefficient, used either to reduce the amount of calculation or to make the coefficient less sensitive to non-normality in distributions [21, 22].

IV. COMPARISON BETWEEN REGIONS

As we have stated, the purpose of this study is to compare the topic categories between regions in order to understand the regional similarity and thus help make better use of Twitter as an information source to find leading trends. In this section we compare and analyze the full set of data as described in section II. To do this work, we classify each of the followers into 8 categories by their acting region as shown in Table 3. Table 3 shows the sum of the followers for 8 divided categories. The category Arts is commonly strong among all of the investigated regions. Particularly, Arts is very high in Global but relatively lower in Tokyo.

Table 3. Comparison among Global, Seoul, Tokyo, and Beijing

	Arts	Business	Education	Government	Life	News	Sci&Tech	Twitter
Global	74	9	0	2	4	9	0	2
Seoul	44	13	4	5	30	1	3	0
Tokyo	27	27	0	7	28	5	5	1
Beijing	30	17	6	4	25	9	8	1

A. Distribution of Topic Categories by Global, Seoul, Tokyo, and Beijing Groups

Many of the top 100 Twitter users are concentrated in a single category, as shown in Fig. 4. The most famous users in Twitter almost all belong to the Arts category. The Arts

is a strong category in Twitter. In the Arts category in each of Seoul, Tokyo, and Beijing, the number of top users is almost same. On the other hand, as can be seen in Fig. 5, many famous top 100 users are classified into the Life category in Seoul, Tokyo, and Beijing. Fig. 6 presents the distribution of categories of the top 100 Twitter users by region. The statistics in Fig. 6 include the mean, median, and standard deviation for 8 topic categories. The vertical bars in Fig. 6 show standard deviations. In the Arts and Life categories, the standard deviation is larger than in any of the other categories. In the Arts category, this is due to the larger number of top 100 users from the Global group relative to the other groups, and in the Life category, it is due to the relatively small total number of users classified into this category.

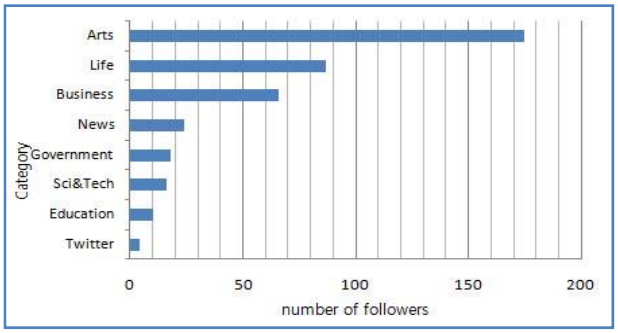

Fig. 4. Distribution of sum size of top 100 Twitter users for each category.

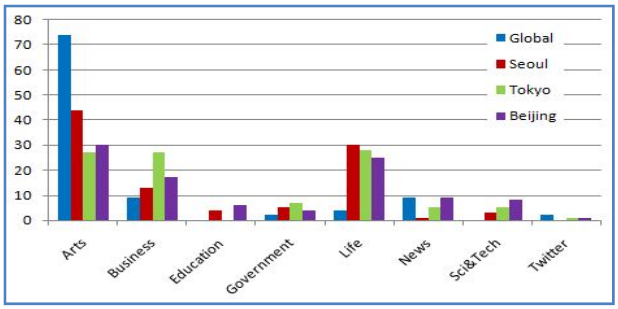

Fig. 5. Comparison of number of top 100 Twitter users in each category for the Global and East Asia groups.

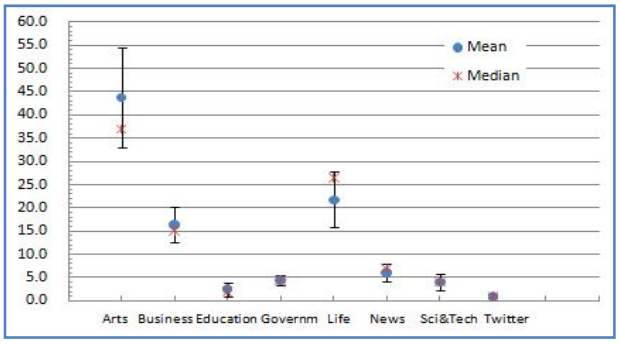

Fig. 6. Distribution of top 100 Twitter users by topic categories in the Global and East Asia groups.

B. Comparison of Topic Categories by East Asia: Seoul, Tokyo, and Beijing

We are interested in finding categorical similarities among Seoul, Tokyo, and Beijing. To observe this in detail, Fig. 7 shows the number of top 100 users for the different categories in Seoul, Tokyo, and Beijing, but not the Global group. Here we can see some of the differences in the Arts category: Seoul is a little higher than the other two regions. This means that the number of followers in Seoul is larger than that in Tokyo and Beijing with regard to the arts. However, when compared to the Global representation in the Arts category, the representation in Seoul is not significant. These differences need to be analyzed in greater detail by using a correlation coefficient to measure the regional similarity. As can be seen in Fig. 8, the total number of standard deviations is decreased and each category becomes sticky. We found significant similarities in the values of the standard deviation for topic categories. When we limit the analysis to East Asia, regional similarity is revealed according to topic categories as shown in Fig. 8. This means that our statistical analyses are performed only for East Asia.

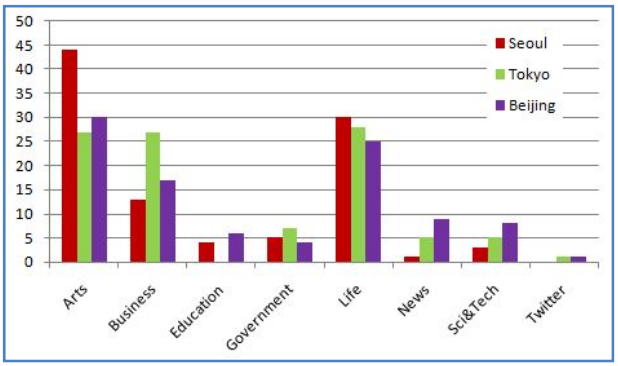

Fig. 7. Comparison number of top 100 Twitter users in each category for East Asia.

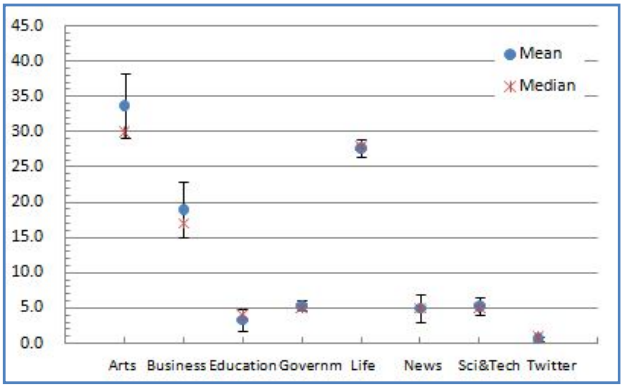

Fig. 8. Distribution of top 100 Twitter users with regions by topic categories in East Asia.

C. Correlation Analysis between the Global, Seoul, Tokyo, and Beijing Groups

Table 4 shows the correlation coefficients between the Global, Seoul, Tokyo, and Beijing groups; these values are calculated by the definition of the Spearman correlation coefficient. The correlation coefficient between the Global and Seoul groups is $\rho = 0.51$; this value is the lowest one among the six values in Table 4. However, this result means that the regional similarity for the categories is comparatively strong with positive correlations between the Global and Seoul groups. The Global and Tokyo groups havea similar correlation coefficient as can be seen from the Table 4. These two cases have similar trends in topic categories.

Table 4. The correlation between Global, Seoul, Tokyo, and Beijing

	Global	Seoul	Tokyo
Seoul	0.51	-	-
Tokyo	0.59	0.82	-
Beijing	0.72	0.76	0.74

The highest value in Table 4 is $\rho = 0.82$ between Seoul and Tokyo. This value shows that the regional similarity for the categories is a very strong positive correlation between Seoul and Tokyo. The result can be interpreted as showing that these two regions have almost the same topic categories. Therefore, the similarity between these two regions is very high. The values $\rho = 0.72$, $\rho = 0.76$, and $\rho = 0.74$ represent astrong positive correlation. We can see these values on the third row in Table 4. The regional similarities for the categories between Beijing and Global, Beijing and Seoul, and Beijing and Tokyo are all high.

We can say that the regional similarity for categories is high among East Asian countries based on the statistical results. Particularly, the correlation between Seoul and Tokyo is very high for topic categories on Twitter. The correlation between Global and East Asia is relatively lower than the correlation among the three East Asian groups.

V. CONCLUSIONS

Twitter is a micro-blogging service that has quickly organized a social networking service that can be used to obtain facts, opinions, and ideas, and even generate collectiveness in real time. The popularity of this social media has attracted the attention of many researchers. Among these studies, some have carried out content analysis of Twitter.

In this paper, we empirically compared and analyzed the categorical trends of Twitter with all of the top 100 users in each of four geographic regions. We thoroughly classified the specific topics into broader categories and analyzed their characteristics with statistical methods. To analyze the categories, we mainly considered their regional relationship and selected four regions: Global, Seoul, Tokyo, and Beijing. In the preliminary investigation, the number of followers of the top 10 Twitter users differed greatly by regional group, with a large difference between the US-centric Global group and the East Asian groups. The Seoul, Tokyo, and Beijing groups, but not the Global group, had a similar number of followers of their top 10 Twitter users. Our purpose was the analysis of the top 100 Twitter users in each of the regions. Based on the overall sums for each topic, the Arts category was the largest and the second ranking category was Life. In this case, the standard deviation between the Global group and East Asian groups was slightly larger than those among the other categories. When we observed only East Asia in more detail, then the standard deviations decreased; therefore, each category become similar. The correlation between the Global and Seoul groups was the lowest value among the six relationships. However, it was still statistically significant. This study found that Seoul, Tokyo, and Beijing, all of East Asia, have a high regional similarity in terms of topic categories in Twitter. Based on the correlation analysis, Seoul and Tokyo showed a particularly sticky trend. In the statistics, the correlation coefficient showed a very strong positive correlation between Seoul and Tokyo. The correlation between the Global group and the groups of East Asia was relatively lower than the correlation among the East Asian groups.

REFERENCES

[1] D. Zhao and M. B. Rosson, “How and why people Twitter: the role that micro-blogging plays in informal communication at work,” *Proceedings of the ACM 2009 International Conference on Supporting Group Work*, New York: NY, pp. 243-252, 2009.

[2] A. Java, X. Song, T. Finin, and B. Tseng, “Why we Twitter: an analysis of a microblogging community,” *Advances in Web Mining and Web Usage Analysis (Lecture Notes in Computer Science, vol. 5439)*, pp. 118-138, 2009.

[3] C. Honeycutt and S. Herring, “Beyond microblogging: conversation and collaboration via Twitter,” *Proceedings of the 42nd Hawaii International Conference on System Sciences*, Big Island: HI, pp. 1-10, 2009.

[4] J. Dixon and R. C. Tucker, “We use technology, but do we use technology? Using existing technologies to communicate, collaborate, and provide support,” *Proceedings of the 37th Annual ACM SIGUCCS Fall Conference on User Services Conference*, New York: NY, pp. 309-312, 2009.

[5] B. McNely, “Backchannel persistence and collaborative meaning-

making," *Proceedings of the 27th ACM International Conference on Design of Communication*, Bloomington: IN, pp. 297-304, 2009.

[6] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shake Twitter users: real-time event detection by social sensors," *Proceedings of 19th annual International Conference on World Wide Web*, Raleigh: NC, pp. 851-860, 2010.

[7] K. Starbird, L. Palen, A. L. Hughes, and S. Vieweg, "Chatter on the red: what hazards threat reveals about the social life of microblogged information," *Proceedings of the ACM 2010 Conference on Computer Supported Cooperative Work*, New York: NY, pp. 241-250, 2010.

[8] A. L. Hughes and L. Palen, "Twitter adoption and use in mass convergence and emergency events," *Proceedings of the 6th International Information Systems for Crisis Response and Management Conference*, Gothenburg, Sweden, 2009.

[9] S. Vieweg, L. Palen, S. B. Liu, A. Hughes, and J. Sutton, "Collective intelligence in disaster: an examination of the phenomenon in the aftermath of the 2007 Virginia Tech Shootings," *Proceedings of the 5th International Information Systems for Crisis Response and Management Conference*, Washington: DC, 2008.

[10] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li, "Comparing Twitter and traditional media using topic model," *Advances in Information Retrieval (Lecture Note in Computer Science, vol. 6611)*, pp. 338-349, 2011.

[11] 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 2010 IEEE Second International Conference on Social Computing*, Minneapolis: MN, pp. 177-184, 2010.

[12] 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, pp. 65-74, 2011.

[13] R. D. Waters and J. Y. Jamal, "Tweet, tweet, tweet: a content analysis of nonprofit organizations' Twitter updates," *Public Relations Review*, vol. 37, no. 3, pp. 321-324, 2011.

[14] 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.

[15] J. J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi, "Short and tweet: experiments on recommending content from information stream," *Proceedings of the 28th International Conference on Human Factors in Computing System*, Atlanta: GA, pp. 1185-1194, 2010.

[16] H. Kwak, H. Chun, and S. Moon, "Fragile online relationship: a first look at unfollow dynamics in Twitter," *Proceedings of the 29th International Conference on Human Factors in Computing System*, Vancouver: BC, pp. 1091-1100, 2011.

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

[18] 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.

[19] Twiter, Inc., Twitter Counter [Internet]. Available: http://twittercounter.com.

[20] Nathan Reed, Gigatweeter [Internet]. Available: http://gigatweeter.com/analytics.

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

[22] J. S. Maritz, *Distribution-free Statistical Methods*, New York, NY: Chapman & Hall, 1981.

Hong-won Yun

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