Classifying Temporal Topics with Similar Patterns on Twitter

Hongwon Yun, *Member, KIMICS*

***Abstract*— Twitter is a popular microblogging service that enables the users to send and read short text messages. These messages are becoming source to analyze topic trends and identify relations among temporal topics. In this paper, we propose a method to classify the temporal topics on Twitter as a problem of grouping the similar patterns. To provide a starting point for a classification under the same topics, we identify the content word weighting scheme based on Latent Dirichlet Allocation (LDA). And we formulate how the temporal topics in the time window can be classified like peaky topics, constant topics, and periodic topics. We provide different real case studies which show the validity of the proposed method. Evaluations show that the proposed method is useful as a classifying model in the analysis of the temporal topics**

***Index Terms*— Temporal topics, Peaky topics, Constant topics, Periodic topics, Twitter.**

I. INTRODUCTION

TWITTER, a popular microblogging service, is a variation on blogging in which users can write short messages to a new form blog that are subsequently distributed to their followers and other observers. One of the popular microblogging services is Twitter that enables to broadcast and share information about their thoughts, opinions, and activities [1]. Users use microblogging to talk about their activities and to find and share information. Twitter, on which we focus here, allows its users to send messages which retrieved through a wide variety of clients. Twitter enables real-time distribution of information to a large group of users directly. According to TechCrunch, Twitter is now attracting 190 million visitors per month and generating 65 million Tweets a day. These numbers are up slightly from 180 million self-reported unique visitors per month back in April, and 50 million Tweets per day in February 2010 [2]. The long term average of tweets is given in Figure 1, we can see that the average tweets per second in October 2010 has been increasing more three times than it in November 2009 [13].

Java et al. identified several categories of intention to use Twitter, including daily chatter, sharing information of URLs, reporting news and conversation [1]. Researchers have been seeking how to use these Twitter messages. There are some results that this potential is already being realized. Twitter is being used to disseminate information in institutional setting and to connect groups of people in critical situations. Twitter has the potential to be used for sharing and coordinating activities [3]. An important common characteristic in Twitter is its real-time nature. The large number of tweets results in numerous reports to events. Thus, the analysis of Twitter data enables to detect the emergent events like earthquake, wildfire, heavy rain and storm. Some results related to them have been accomplished [4]-[6].

Recently, to organize the Twitter messages as conversations based on their semantic meaning has received much attention [7]. How to find topics and trends within big amount of Twitter stream is a challenging problem. As a similar example, the emerging terms during fifty days about “Human Interest” is shown in Figure 2 [14]. Figure 3 shows that the top fifteen trends during thirty days until March 31, 2011. Each row contains the 15 top ranking words and each column corresponds to a date [14]. Twitter data are distributed as stream over time, providing an ongoing commentary of topics, trends, and issues [8]. Our interest in the use of Twitter data is finding temporal topics. We wish to examine the ongoing current topics within these Twitter data such as peaky topics, consistent issues and regular conversation. A significant task of this analyzing Twitter messages is to observe and track the popular events, or topics that evolve over time in the social network, specifically on Twitter.

In this paper, our purpose is to find momentary topics with peaks and what is being talked among Twitter users during relatively long time in the background and what topics are discussed regularly over time.

The rest of this paper is organized as follows. In Section 2, we basically classify the temporal topics change over time on Twitter and explain the characteristic of Twitter data. In Section 3, we describe a content word weighting scheme as a novel topic model for Twitter data based on Latent Dirichlet Allocation. And we present the formula to classify temporal topics with similar shapes. Evaluation on real case study is given in Section 4. Finally, we summarize our research in Section 5.

Manuscript received May 6, 2011; revised June 2, 2011; accepted June 7, 2011.

Hongwon Yun is with the Department of IT, Silla University, Busan, 609-736, Korea (Email: hwyun@silla.ac.kr)

II. TEMPORAL TOPICS ON TWITTER

A. Classification of Temporal Topics

A huge volume of Twitter data is generated from the social communities such as blogs, microblogs. Both the conversations and the contents on these social communities are changing over time. The Twitter messages have a great deal about the structure and trend of the event contained in the stream [9].

We believe that important moments will have words closely related with them that are extremely frequent in the time window, whereas they might be relatively infrequent at other times. The words associated with peaky topics are particular to an exact the time window and not noticeable to other time windows. For example, a massive 9.0 magnitude quake hit Northeastern Japan on Friday, March 11 2011, causing devastating flooding from 10 meter-high tsunamis that hit the Eastern coastline. As expected, the devastation in Japan from earthquake and tsunami captured the thought of Twitter users from around the world and made this topic number one [16].

We also expect to find constant topics which are less salient words and topics than peaky topics. They are continued particularly for time duration. We guess that persistent topics with continued levels of conversation would be reflected Twitter user's interest over temporal evolution. We also wish to find regular conversations. There are repeatedly appeared issues that people have an interest in their living. The text of tweets can reveal a great deal about the activity of events regularly. We expect that moments of interests will have words associated them that are regularly repetition in the time windows.

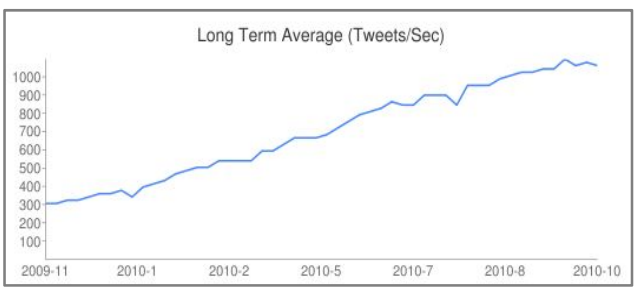

Fig. 1. Long term average of tweets

Fig. 2. The emerging terms during fifty days about "Human Interest"

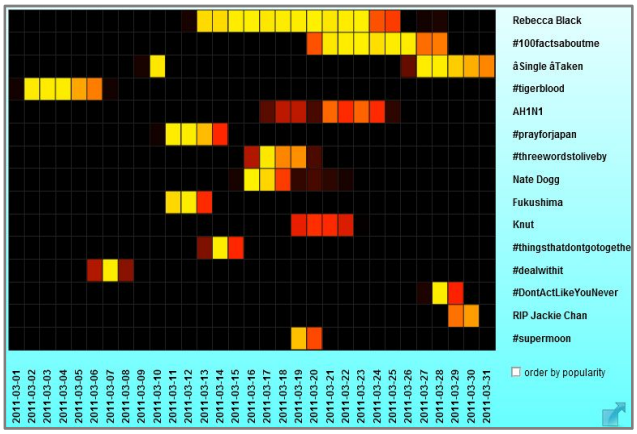

Fig. 3. The top 15 trends during 30 days until March 31, 2011.

B. Consideration of Tweets Characteristics

Twitter data are short, heterogeneous and noisy. Twitter messages often lack proper written grammar and punctuation. They frequently contain abbreviation and internet slang. Traditional natural language processing methods are often inadequate for processing this type of data [10]. There are low content words and short sentences in Twitter messages. These terms are significant part of each message and significantly influence their expression. A couple of words in tweets carry more semantic content than other documents. The short messages contain extremely little lexical redundancy. We must consider these content words with higher informative value. It deserves to be weighted more heavily. A Twitter messages can deal with a couple of topics, and the terms that appear in that tweet reflect the particular set of interests. The fact that one or two words can be responsible for the topics occurring in a single Twitter message discriminates this application of the topics model from the topics models in traditional natural language processing.

III. CLASSIFYING TEMPORAL TOPICS

A. Content Word Weighting Scheme

We have explored the use of topics models to analyze Twitter data. Considering that the Twitter messages are usually short, sparse, and do not include statistical repetition. It is quite dissimilar in large document. Our approach we use in this paper is based on latent variable topic model like LDA [11]. LDA is a generative probabilistic model for collection of discrete data such as text corpora. The model finds latent structure in a collection of documents and can be used to discover topics in a document.

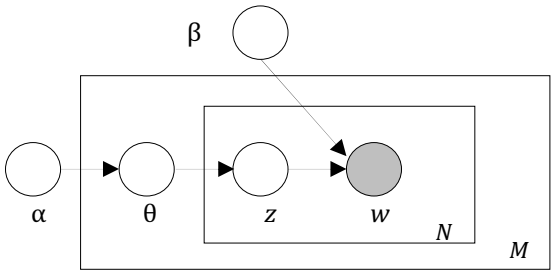

Fig. 4. Graphical model representation of LDA

The LDA model is represented as a probabilistic graphical model in Figure 4. According to [11], the parameter α and β are corpus level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ are document level variables, sampled once per document. The variables z and w are word level variables are sampled once for each word in each document [11].

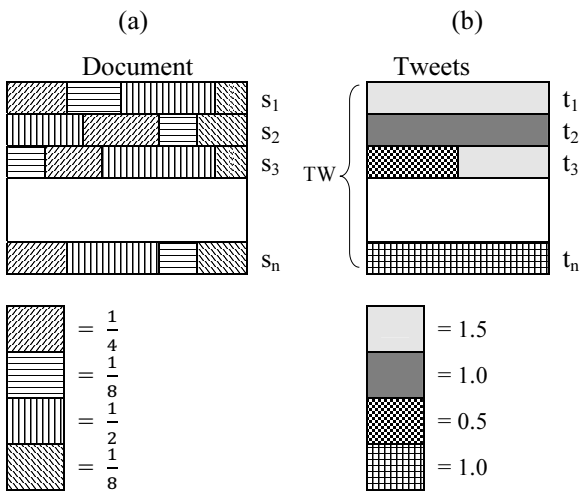

Fig. 5. Comparison of probabilities of individual topic in a document and tweets

In the topic model, words in a document are represented as belonging to set of probabilistic topics. The figure 5a shown on the above rectangle represents a document with four topics and s_1, s_2, … s_n are sentences compose a document. The below of figure 2a indicates the probability of each topics under four topics in a document. And the sum of the probability of each topic in a document is 1. Figure 5b shows *n* or more topics in a time window and t_1, t_2, … , t_n are tweets. TW on the figure 5b means time window which is determined by user. To search for emerging topics on Twitter, the analysis processes begins with the real time extraction from the stream of tweets. A time interval is to search the relevant words as $T^i = (t_i, t_i + d)$ where t_i is the starting point of the i*th* time interval. We extract the corpus W^i, and a number of the corpus extracted during the time interval T^i. Each tweet has probability less than or equal to 1.0. Therefore the sum of the probability of topic in a time window is greater than or equal to 1.0. Since a tweet usually one or two topics and each tweet is separated from other one.

As mentioned previous, Twitter data are composed of short content words. We can focus on only one θ from the original LDA model [11] as following.

$$\int_{\theta j} P(\theta_j;\alpha) \prod_{t=1}^{N} P(Z_{j,t}|\theta_j) d\theta_j$$

It is the hidden part of the model the j^{th} document. The above equation can be changed as following.

$$\frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{K}\Gamma(\alpha_i)} \frac{\sum_{i=1}^{K}\Gamma(n_{j,(\cdot)}^{i}+\alpha_i)}{\Gamma(\sum_{i=1}^{K} n_{j,(\cdot)}^{i}+\alpha_i)}$$

We apply the new word weighting scheme instead of the original θ to give content words more weight. Content words in the twitter message have a higher value when θ is computed for topic assignments. The content word weighting scheme is based on the tf-idf weight.

$$(tf-idf)_{i,j} = tf_{i,j} \times idf_i$$

where $tf_{i,j}$ is the term frequency and idf_i is the inverse document frequency [12]. This content word weighting scheme can be used to impose weight a specific word in the tweets.

The topic model using content word weighting scheme (called CWWS topic model) provides a starting point for a classification under the same topics. A tweet can deal with a couple of topics and the words that appear in that Twitter messages reflect the particular topics. All twitter messages are regarded to be classified under topics by the CWWS topic model. Therefore, we can assume that each tweet in the time window is grouped into an individual topic.

B. Metrics for Classifying the Similar Patterns

As mentioned in the introduction, we wish to group temporal topics on Twitter in a given time window. Thus, we define the i*th* considered a time window T^i as

$$T^i = < t_i, t_i + r >$$

where t_i is the starting instant of the i*th* considered time window. This given a time range set by user. Twitter messages are extracted and grouped during the time window T^i. To state conveniently, the twitter data referred below already have been grouped into a specified sub collection. We suppose the twitter messages in the sub collection have been gathered in the time window T^i. Here, the whole twitter messages collection W^i in T^i could be represented as:

$$W^i = \{w_1^i, w_2^i, \dots, w_i^i\}$$

Each element in W^i is grouped into an individual topic. And there are totally *n* twitter messages in the whole messages. Each topic within the time range has a number of messages such as $m_1^i, m_2^i, \dots, m_i^i$. Thus, given

the time interval, we calculate the average frequencies for the individual topic and define them as below:

$$F^i = \{f_1^i, f_2^i, \dots, f_i^i\}$$

We need to distinguish between peaky topics, constant topics and regularly repeated topics. There may be many ways to describe the difference between three topics using statistical methods. An alternative is to use the standard deviation. Given the same interval, the standard deviations for the individual topic are defined as following:

$$D^i = \{d_1^i, d_2^i, \dots, d_i^i\}$$

For specific time window, a local maximum (called max) and a local minimum (called min) at the frequencies can be found. If there exists some $\delta > 0$ such that $|lmax - f_i^i| \cong \delta$ and $|lmin - f_i^i| \cong \delta$ when $|lmax - lmin| \geq 0$, the value δ can be useful to determine for periodic topics rather than peaky and constant topics.

The average frequency and standard deviation of individual topic can be generalized as *f* and *d* respectively. We use *f* and *d* in order to differentiate between peaky topics, constant topics, or periodic topics. We define the formula $f - d < 0$ and $d/f > n$ as the criteria for peaky topics, where *n* is the integer, In fact, the word identified as peaky like emergent events only if its value of d/f is *n* times. We define the formula $f - d > 0$ and $d/f < 1$ as the criteria for constant topics. This measure tends to give low values. The values calculated from x/f can be used as auxiliary values to differentiate between peaky, constant, or periodic topics. The symbol x means that the maximum frequency within a given time window. The highest value from x/f tends to indicate an emergent event, on the contrary, the lowest value generally represents conversational topics. Using these criteria we are able to distinguish the topics into similar patterns.

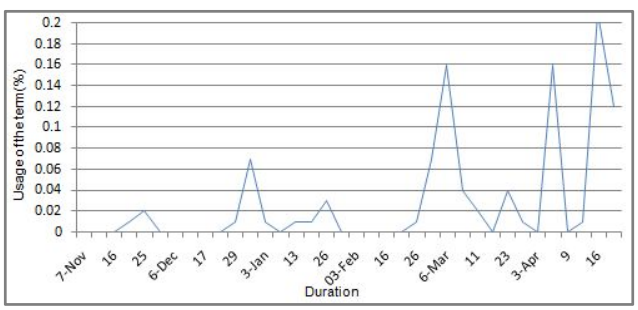

Fig. 6. An example of users' activity data on Twitter

To show how Twitter was used during the tornadoes in USA, we collected and sampled users' activity data during the time window between November 2010 and April 2011. In Figure 6, we show the user's activity for "Tornado" and statistical values as an example.

IV. EVALUATIONS ON REAL CASE STUDY

A. Data Collection

In this section, we evaluate our proposed method by analyzing real case studies. Twitter is a popular microblogging service that enables the users to send and read short text messages known as tweets. In particular, we collected a sample of tweets on Trendistic [15] during the period included between 31th of October 2010 and 20th of April 2011. The original charts were provided by Trendistic.

We wish to classify a Twitter message into two or more mutually exclusive topics and each temporal topic is classified into one or more specific topics. In the first step, we can apply the topic model using CWWS to classify the topics for Twitter messages. We assume that Twitter messages in the specified time interval already are classified into proper topics through the first step. We analyze all of the real case studies using the formula as the criteria for individual temporal topics. Five visualized example are given to compare our experimental results with real-life events.

B. Analysis on Real Case Study

We consider the time interval related to the keywords as mentioned previous and show the five words and their statistical values in Table 1. We note, as this is early work, this analyzing values offer an interesting task to collect and analyze much more experimental data set continuously. In Table 1, the statistical values of each keyword used to discriminate the topic patterns such as peaky topics, constant topics, or periodic topics.

TABLE 1
THE VALUE COMPUTATION ON THE KEYWORDS

Stat / Term	Avg	Stdev	Max	Min	Aux
Tsunami	0.32	1.04	6.24	0.00	19.5
Tornado	0.03	0.05	0.21	0.00	7.0
Food	0.32	0.08	0.63	0.22	1.97
Love	2.54	0.39	3.77	2.04	1.48
Friday	0.42	0.33	1.65	0.08	3.93

Statistical values and usage of the word "Tsunami" in Twitter from October 2010 to April 2011 is shown in Figure 7. The peak shows the catastrophic tsunami occurred in Japan on 11 March 2011. Words like "Tsunami" have very peaky usages in the Twitter community due to the fact that they represent emergent events. Accordingly, we can see that $\bar{x} - \sigma < 0$ and $\bar{m}/\bar{x} = 19.5$ in Figure 7 when we apply the formula as described in Section III.B.

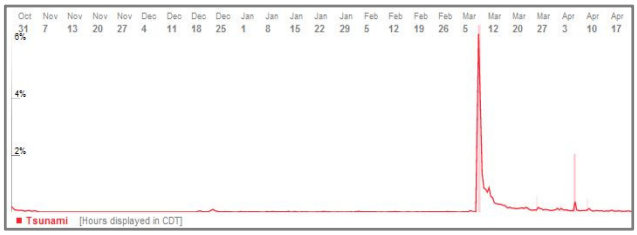

Fig. 7. Statistics based on tweet frequency of the word "Tsunami"

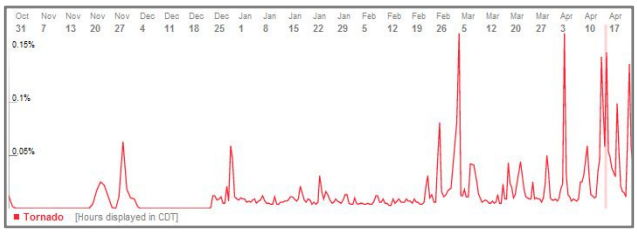

Fig. 8. Statistics based on tweet frequency of the word "Tornado"

Another interesting example we find in Twitter topics is about tornadoes. During the spring and early summer, thousands of tornadoes strike the United States. Across the United States, the April 16 tornadoes and storms killed 24 people. In Figure 8, the highest peaky shapes represents that deadly tornadoes were occurred. As explained in the previous paragraph, it could be classified as peaky topic after analyzing the statistical values on the topics.

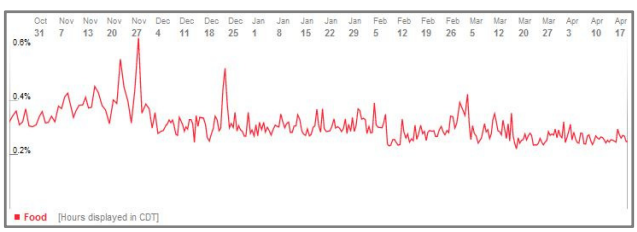

Fig. 9. Statistics based on tweet frequency of the word "Food"

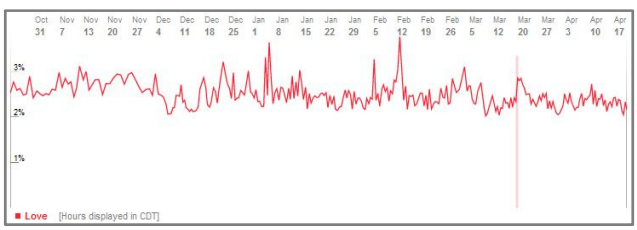

Fig. 10. Statistics based on tweet frequency of the word "Love"

Terms like "Food" and "Love" have very constant and standard usage in the community due to the fact that Twitter users represent common and necessary activities. As shown in Figure 9 and 10, the criteria for constant topics $f - d > 0$ and $d/f < 1$ is satisfied as well. We found that a low standard deviation is a proper indicator that the word is used for conversational in the Twitter community.

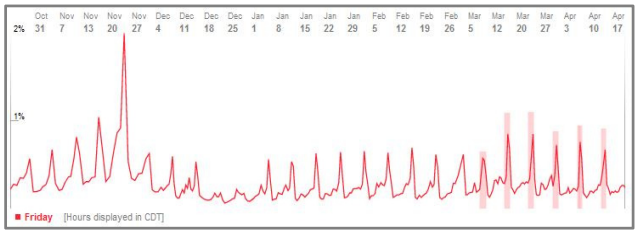

Fig. 11. Statistics based on tweet frequency of the word "Friday"

Keyword "Friday" has higher usages on Friday and lower usage during Saturday and Thursday. These words like "Friday" have tweets cycle in the Twitter due to the fact that they mean periodic events. In order to indentify how the words have been classified as periodic topics, we need to consider their related periods such as $\delta = 7$ in Figure 11.

V. CONCLUSIONS

Twitter offers a social networking and microblogging service. Twitter enables its users to send and read messages. The users can group posts together by topic or type by use of hashtags. This popular microblogging service has been used for a variety of purpose in many different industries and scenarios. The text of tweets contains a great deal of information like topics, trends, and issues. We wished to examine the ongoing current topics within these Twitter messages like peaky topics, constant issues, and periodic topics.

We have explored the use of topics models to analyze Twitter data. Thus, we identified the content word weighting scheme to impose weight a specific word in the tweets. This scheme provides a starting point for a classification under the same topics. Also, we formulated the task of classifying the temporal topics on Twitter as a problem of grouping the similar patterns within specified time interval. Therefore, we were able to classify the groups of temporal topics which have similar type and level of interest over time. We provided different real case studies which show the validity of the proposed method. From the evaluation results, the proposed method is useful as a classifying model in the analysis of the temporal topics.

REFERENCES

[1] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: An Analysis of a Microblogging Community," *Proc. the 9th WebKDD and 1st SNA-KDD 2007*, LNCS 5439, pp. 118-138, 2009.

[2] TechCrunch, Available: http://techcrunch.com/2010/06/08/twitter-190-million-users
[3] C. Honeycutt and S. C. Herring, "Beyond Microblogging: Conversation and Collaboration via Twitter," *Proc. the 42nd Hawaii International Conference on System Sciences 2009*, IEEE Press, 2009.
[4] T. Sakaki, M. Okazaki, Y. Matsuo, "Earthquake Shake Twiiter Users: Real-time Event Detection by Social Sensors," *Proc. WWW 2010*, pp. 851-860, April, 2010.
[5] M. guy, P. Earle, C. Ostrum, K. Gruchalla, and S. Horvath, "Integration and Dissemination of Citizen Reported and Seismically Derived Earthquake Information via Social Network Technologies," *Proc. IDA 2010*, pp. 42-53, 2010.
[6] A. Hughes, L. Palen, "Twitter Adoption and Use in Mass Convergence and Emergency Events," *Proc. the 2009 Information Systems for Crisis Response and Management Conference*, 2009.
[7] Ye Tian et al. "Topic Detection and Organization of Mobile Text Messages," *Proc. CIKM'10*, pp.1877-1880, October, 2010.
[8] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Peaks and Persistence: Modeling the Shape of Microblog Conversations," Proc. CSCW 2011, March, 2011.
[9] C. X. Lin, B. Zhao, Q. Mei, and J. Han, "PET: A Statistical Model for Popular Events Tracking in Social Communities," *Proc. KDD'10*, pp. 929-938, July, 2010.
[10] K. Kireyev, L. Palen, K. M. Anderson, "Application of Topics Models to Analysis of Disaster-Related Twitter Data," *In NIPS Workshop on Applications for Topic Models: Text and Beyond*, December, 2009.
[11] D. M. Blei, A. Y. Ng, and M. I. Jodan, " Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3, pp. 993-1022, 2003.
[12] K. Sparck-Jones, "A Statistical Interpretation of Term Specificity and its Application in Retrieval," *Journal of Documentaion*, Vol. 28, No. 1, 1973.
[13] Long term average of tweet. Available: http://gigatweeter.com/analytics
[14] Trends on Twitter aggregator: Available: http://twopular.com
[15] Trends in Twitter: Available: http://trendistic.com
[16] Top 10 Twitter trends: Available: http://mashable.com
[17] Twitter API: Available: http://apiwiki.twitter.com
[18] Real time Twitter trends: Available:http://tweettabs.com

Hongwon Yun
He received his B.S. and the Ph.D. degrees at the Department of Computer Science from 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 database, temporal database and social network.