2013 International Conference on Social Intelligence and Technology

RT^{2}M : Real-time Twitter Trend Mining System

Min Song
Dept. of Library and Information Science
Yonsei University
Seoul, Republic of Korea
min.song@yonsei.ac.kr

Meen Chul Kim
Dept. of Library and Information Science
Yonsei University
Seoul, Republic of Korea
andrewevans@yonsei.ac.kr

Abstract—**The advent of social media is changing the existing information behavior by letting users access to real-time online information channels without the constraints of time and space. It also generates a huge amount of data worth discovering novel knowledge. Social media, therefore, has created an enormous challenge for scientists trying to keep pace with developments in their field. Most of the previous studies have adopted broad-brush approaches which tend to result in providing limited analysis. To handle these problems properly, we introduce our real-time Twitter trend mining system, RT²M, which operates in real-time to process big stream datasets available on Twitter. The system offers the functions of term co-occurrence retrieval, visualization of Twitter users by query, similarity calculation between two users, Topic Modeling to keep track of changes of topical trend, and analysis on mention-based user networks. We also demonstrate an empirical study on 2012 Korean presidential election. The case study reveals Twitter could be a useful source to detect and predict the advent and changes of social issues, and analysis of mention-based user networks could show different aspects of user behaviors.**

Keywords—social media mining; real-time Twitter trend miming system; topic modeling; network analysis; Korean presidential election

I. INTRODUCTION

Social media is a representative form of the Web 2.0 services which has undoubtedly contributed to the emergence of Big Data, based on voluntary and active participation of users. Throughout the world, social media receives spotlights by a lot of Web users because of the benefits in that 1) no professional knowledge and skills are required to produce and share contents in real-time and 2) social network among users with similar interests can be easily formed [2]. Facebook and Twitter, representative social media services, have estimated users of each 850 million and 500 million people all over the world [30].

On the other hand, a variety of research areas such as social network analysis, opinion mining, and so on has paid attention to extract the meaningful information among vast amounts of data daily produced in social media. For example, the presidential election camp which led U.S. President Barack Obama to be successfully re-elected recruited many data experts for the analysis of social media data while preparing the last presidential election. Like this instance, social media has been main foci to the field of Information Retrieval and Text Mining from its first advent because not only it produces massive unstructured textual data and user relations in real-time but also it could be used as an influential channel for opinion leading such as agenda-setting and public opinion formation. This creates an enormous challenge for scientists trying to keep pace with developments in their field.

Until now, however, most of the previous studies have focused on somewhat limited approaches including 1) retroactively designed experimental conditions which give dynamic queries to static data collection [6][12][15][22], 2) generic issue modeling [6][15][30], and 3) social network analysis based on simply visible connectivity such as follow/following relationship in Twitter [3][13]. These approaches could face with following daunting challenges in finding and analyzing new information. First, enormous social media data is being generated in real-time. It could be, therefore, ineffective for using static data to detect continuously changing social trends such as users' interests and public issues. Second, a specific issue is likely to be composed of several subordinate events. Thus, it needs to examine formation and extinction of the related events, and users' reaction for effective analysis of the issue from various angles. Finally, user network in Twitter has different characteristics from those of other social media. In other words, users in Twitter can make relations with others by sending-receiving mentions. Hence, only concentrating on the follow/following connection cannot properly reflect the trait of the user networks in Twitter. In addition, there is a growing need to identify dynamic changes of user networks accompanied by those of social issues in the real world.

Therefore, our study aims at designing and developing the RT^{2}M (Real-time Twitter Trend Mining) system which allows in real-time to 1) crawl and store every textual data ("tweets") generated in Twitter into a local database, 2) keep track of social issues by temporal Topic Modeling, and 3) visualize mention-based user networks. We also demonstrate a case study related to 2012 Korean presidential election carried by the RT²M system.

The rest of the article is organized as follows: Section 2 surveys on related works. Section 3 provides the detailed descriptions of the proposed methodology. Section 4 reports a case study, and Section 5 concludes the paper.

This study was supported by National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-2012S1A3A2033291).

978-0-7695-4998-9/13 $26.00 © 2013 IEEE
DOI 10.1109/SOCIETY.2013.19

IEEE computer society

II. Related Works

Social media draws every attention from various scholastic communities for it offers a whole new opportunity to understand the social practices. Diverse academic community, therefore, focus on various methodologies such as text mining, network analysis, opinion mining, and etc. with social media to discover unrevealed knowledge and information. Against this backdrop, initial studies on social media focused on attempts to understand user behavior in social media and structure of user community [8][11]. They, however, simply focused on generic connectivity represented by follow/following relations which could not properly take a view of user networks in Twitter because those in Twitter can make relations with others by sending-receiving mentions without visible connection. In addition, other studies revealed that follow/following connection could not be a useful indicator to find an influencer [3][13], and contents and sending-receiving messages should be considered to properly calculate distance and similarity among users [14][22]. We, therefore, aim at analyzing mentioned-based user networks to tackle this problem in a proper manner. In other words, we concentrate on identifying thematic coherence among users resided in the action of sending-receiving mentions. Meanwhile, research communities employ various approaches for mining unrevealed information from social data. Among them, LDA is of great interest in the field of Text Mining for modeling latent issues from unstructured textual data such as the Web and social media [5][30]. Contrary to its sound mathematical foundation, however, it has a fatal limit that there does not exist any validated method for evaluation. In this study, we hire temporal LDA to compare and analyze relationship between topics extracted from tweets on 2012 Korean presidential election and related events in the real world. At the same time, like a study that modeled topics based on term co-occurrence [6], our real-time twitter trend mining system employs the term co-occurrence retrieval technique to trace chronologically co-occurred terms to compensate the limitation of LDA by suggesting double-layered Topic Modeling.

In addition, , the SCAN [28] algorithm suggests faster techniques than other community detection approaches in identifying hub nodes connecting communities as well as removing outliers that play any significant roles in the network. In reality, a hybrid approach combining Topic Modeling and community detection was devised for identifying relationship between research topics and scholarly communities [29]. Our work also employs modularity and Voltage-clustering algorithm to detect invisible community existing among users.

On the one hand, some studies regarded blog and microblog as forum and opinion channel where users discuss political issues, and evaluated their political potentials [10][26]. Recently, not only researches on how microbloging affects to the actual political events [7][25] but also studies about utilizing social media for predicting the results of the presidential elections [15][18][24] were conducted. Nonetheless, they failed to make valid explanation and appropriate prediction, because they did not consider influential variables such as temporal factor and length of the features [12][15][23]. In the present study, we also demonstrate temporal changes of topical trends in Twitter on 2012 Korean presidential election by using RT^{2}M. Various studies also have tried to explain user's behavior on political issues in social media, and make prediction by mining social data. [24] indicated that supporting for a certain candidate in social media could be an indicator to predict a result of election, and sentiment analysis with social data would properly explain voter's political disposition [18][24]. We, therefore, focus on addressing temporal changes of topical trend on the presidential election with real-time data streaming basis along with mention-based social network analysis.

III. Methodology

In this section, we introduce our Twitter mining system, RT^{2}M, which operates in real-time to process big stream datasets available on Twitter. By utilizing RT^{2}M to mine Twitter data on 2012 Korean Presidential Election, we demonstrate the usefulness of RT^{2}M. For demonstration, we specifically focus on topic trend analysis and network analysis to examine presidential issues embedded in Twitter data.

Figure 1 shows the system overview of RT^{2}M. The detailed system description is provided in Section A.

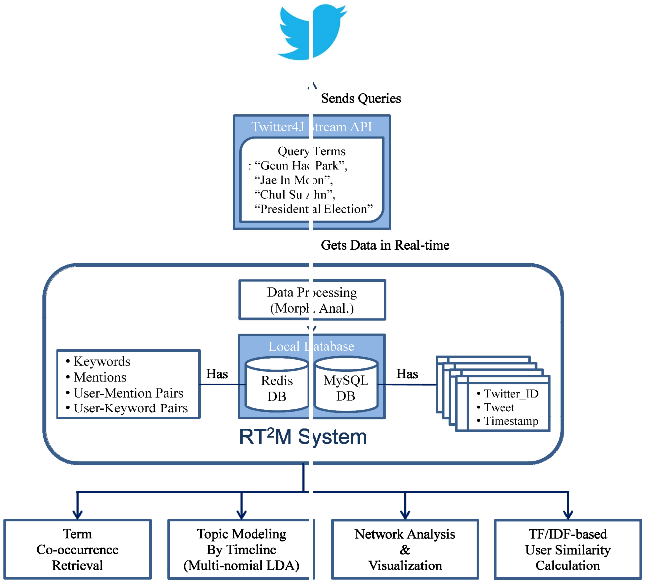


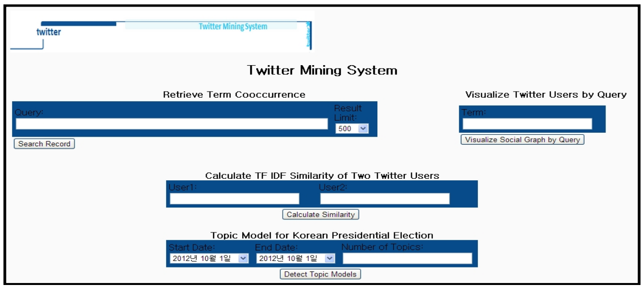
Fig. 1. System Overview of RT^{2}M

A. Twitter Mining System & Data Collection

As aforementioned earlier, the goal of this system is to develop a real-time Twitter trend mining system by effective collection and various text analysis of big stream data. RT^{2}M collects tweets on Twitter in Korea (http://www.twitter.com/), user ID, and time stamps with Twitter Stream API provided in Twitter4J. We store keywords, mentions, as well as pairs of a user and a set of mentions, pairs of a user and a set of keywords in Redis DB. Redis is an open source based key-value store database (http://redis.io). The entire data set is stored in-memory so it is extremely fast. We chose Redis as key-value store because it provides a highly scalable data store shared by multiple processes, multiple applications, and multiple servers.

Along with Redis, we employ the My-SQL relational database to store tweets and time stamps information in disk for further analysis. Redis makes it easier to do the real-time Twitter mining, term co-occurrence retrieval, mention-based Twitter network visualization, and TFIDF-based user similarity. To store keywords, we use the KTS (http://kldp.net/projects/kts/), a probability based Korean morphological analyzer, for spotting important terms. The main reason to use KTS is because it provides sophisticated, well-defined probabilistic rules. However, since tweets contains idiosyncratic terms or jargons, we build a customized lexicon.

The main page of RT²M is shown in Figure 2. Functions provided in RT²M are as follows:



Twitter Mining System

Retrieve Term Cooccurrence

Query:

Result Limit: 500

Search Record

Visualize Twitter Users by Query

Term:

Visualize Social Graph by Query

Calculate TF IDF Similarity of Two Twitter Users

User1:

User2:

Calculate Similarity

Topic Model for Korean Presidential Election

Start Date: 2012년 10월 1일

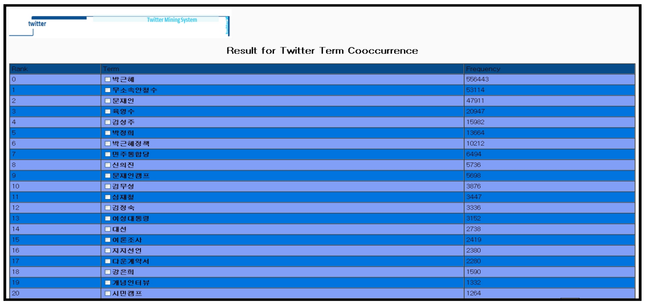
End Date: 2012년 10월 1일

Number of Topics:

Detect Topic Models

Fig. 2. RT^{2}M System

1) Term Co-occurrence Retrieval: Given a query, the system retrieves the list of terms co-occurred with the query term. Once the list is obtained, it sorts co-occurred terms by co-occurrence frequency and can display the result with an option of 100, 500, 1,000, and 2,000 terms. Like other functions provided in the system, co-occurred terms are dynamically updated and displayed as more Twitter stream data is received. Figure 3 below shows one part of the result of term co-occurrence retrieval.



Result for Twitter Term Cooccurrence

Rank	Term	Frequency
0	□박근혜	556443
1	□무소속안철수	53114
2	□문재인	47911
3	□육영수	20947
4	□김성주	15982
5	□박정희	13664
6	□박근혜정책	10212
7	□민주통합당	6494
8	□신의진	5736
9	□문재인캠프	5698
10	□김무성	3876
11	□심재철	3447
12	□김정숙	3336
13	□여성대통령	3152
14	□대선	2738
15	□여론조사	2419
16	□지지선언	2380
17	□다운계약서	2280
18	□강은희	1590
19	□개념인터뷰	1332
20	□시민캠프	1264

Fig. 3. Term Co-occurrence Retrieval with the Query Term "Guen Hae Park"

2) Visualization of Twitter Users by Query: Upon the query term, the system visualizes the social network graph of Twitter users mentioned together with the query term. The visualized graph is bidirectional. The social network analysis for the mention-based Twitter users is provided in the next section. Figure 4 below shows one part of the result of visualizing Twitter users by query.

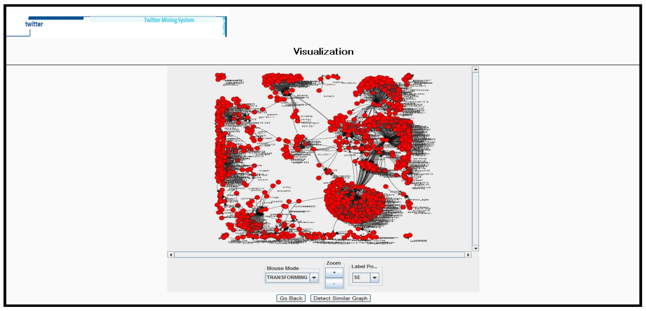


Fig. 4. Visualization of Twitter Users by the Query Term "Jae In Moon"

3) Similarity Calculation between Two Users: It computes cosine similarity between two users comparing terms they use in their tweets, weighted by their tf-idf index. TF-IDF, which stands for term frequency multiplied by inverse document frequency, is a well-received weighting scheme in information retrieval to produce a composite weight for each term in each document [20]. The tf-idf weighting scheme assigns to term t a weight in document d given by $tf \cdot idf_{t,d} = tf_{t,d} * idf_t$. In our work, we count a raw frequency of every term t in one tweet posting, excluding stopwords. Then, we calculate the tf-idf weighting of a certain term, multiplying a tf by an idf dividing the total number of tweets by the number of tweets containing the term, and then taking the logarithm of that quotient.

4) Topic Modeling: The system generates N topics specified by a user along with topical terms for the time period that the user selects. For the topic modeling technique, we employ the Multinomial Latent Dirichlet Allocation technique. The detailed description of the technique is provided in the next section. Figure 5 below shows one part of the result of temporal Topic Modeling. If a start time and end time are entered, the system returns chronological topic trends related to queries which were used upon data crawling.

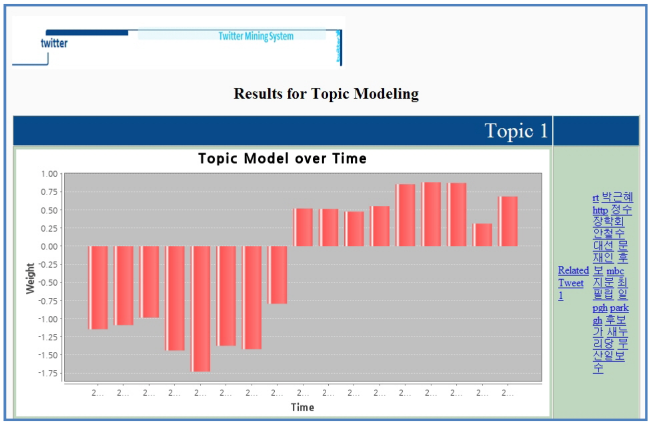


Fig. 5. Multinomial Topic Modeling Result

To examine the performance of RT²M, we collect 1,737,969 tweets with the query term "Geun Hae Park", "Jae In Moon", "Chul Su Ahn", and "Presidential Election" for one month from October 1, 2012 to October 31, 2012.

B. Text Mining Techniques

This section describes text mining techniques that are integrated into RT²M. Some basic mining techniques are explained in Section III.A and henceforth, we describe two important mining techniques of RT²M: Multinomial Topic Modeling and Community Detection for Social Network Analysis.

1) Multinomial Topic Modeling: Document modeling in information retrieval or text mining is a technique that expresses an individual document and the collection of documents in terms of term appearing in documents. Topic modeling is one of the document modeling techniques, and LDA, standing for Latent Dirichlet Allocation proposed by Blei at al. [1], is one of the earliest topic modeling techniques that is based on a graph model with an assumption of Dirichlet prior-based topic distribution. In other words, LDA represents documents as mixtures of topics that spit out words with certain probabilities. The topic modeling technique used in this paper is Dirichlet-multinomial regression (DMR) proposed by Mimno and McCallum [16]. DMR is an extension of Latent Dirichlet Allocation (LDA) proposed by Blei et al. [1], and allows conditioning on arbitrary document features by including a long-linear prior on document-topic distributions that is a function of the features of the document such as author, publication venue, references, and dates. For each document *d*, let x_d be a feature vector representing metadata. Given the prior distribution of $N(0,\Sigma)$and hyper-parameters β, the generative process for documents and their words is as follows:

(1) For each topic *t*, draw $\emptyset_t$~*Dir(β)* noting that *Dir(β)* is a distinct Dirichlet distribution with the Dirichlet prior on the topic-word distribution (a.k.a. hyper-parameters), β

(2) For each document *d*,

a) Draw $\theta_d \sim Dir(\alpha_d) = Dir(exp(\tau_d))\ with\ \tau_d \in \tau$ noting that a per-document α_d, the parameters of a Dirichlet distribution and τ_d is a covariance function $f(y_d, x_k)$ where y_d is the observed attribute vector of document d and x_k is a vector of metadata

b) For each word *w*,

- Draw $Z_{d,w} \sim Multi(\theta_d)$ noting that $Z_{d,w}$ is topic assignment of a word t_w and θ_d is topic proportion of a document *d*
- Draw $T_{d,w} \sim Multi(\emptyset_{Z_{d,w}})$ noting that $T_{d,w}$ is *w*-th word of a document d and $\emptyset_t$ is preference of a topic *t* over the vocabulary with $\sum_n \emptyset_{t,n} = 1$

For DMR topic modeling, we set three fixed parameters: σ^2, the variance of the prior on parameter values for prior distribution; β, the Dirichlet prior on the topic-word distributions; and |T|, the number of topics. The graphical representation of the aforementioned description of a DMR generative model shows in Figure 6.

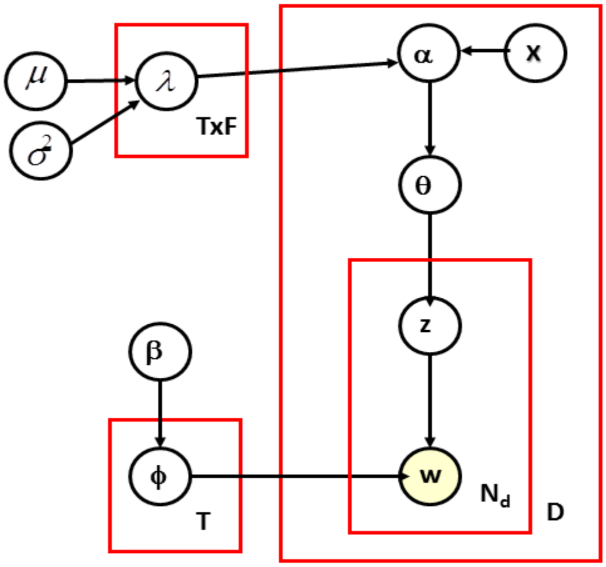


Fig. 6. Graphical Representation of DMR Generative Process Model

2) User Network Analysis: Social Network Analysis (SNA) has been widely used to analyze the relationship of individuals, groups, or societies in terms of network. A person or a group is represented as a node, and each node has the dependent relationship with each other (tie). In social media, since the inter-dependent relationship is the core of the service, extensive studies have been done in the field of SNA. However, the majority of studies on Twitter, a representative conduit of social media, have adopted Twitter's follow/following as the analysis unit. The user network in Twitter has different characteristics from those of other social media. In other words, users in Twitter can make a relationship by acting "follow" with others whom they have interests in their own without the consent of the other party. Hence, only concentrating on the follow/following connection cannot properly reflect the trait of the user networks in Twitter. In addition, dynamic changes of user networks accompanied by those of social issues in the real world need to be identified. To this end, we utilize mentions by Twitter users and their directness for the analysis unit. For network analysis and visualization, we use an open source visualization tool, JUNG standing for Java Universal Network/Graph Framework http://jung.sourceforge.net/), and Voltage-clustering algorithm for detecting user community.

In graph theory, meanwhile, a weak component means a group of nodes, any of which would be reachable from any other node. In our mention-based twitter network, we found that there are 1,537 components consisting of 4,668,343 edges and 136,754 nodes. The communities from large networks carry great scientific and practical value because they typically correspond to behavior or functional units of the network, such as social groups in a social network. A community is a densely connected subset of nodes that is only sparsely linked to the remaining network. Community detection provides us a valuable tool to analyze network structure and better understand complex networks as well as provide better exploration and browsing tools for very large collections. In this paper, we employ the Voltage-clustering algorithm, which is suitable for a large social network. This algorithm is based on a priori that the clusters are of approximately the same size, and uses a more complex method than k-means for determining cluster membership based on co-occurrence data [27].

IV. Case Study: 2012 Korean Presidential Election

In this section, we discuss our case study related to 2012 Korean presidential election carried by the RT²M system. By utilizing RT²M to mine Twitter data on 2012 Korean Presidential Election, we demonstrate the usefulness of RT²M as a real-time text miner. The demonstration is made by comparison of term co-occurrence by query, topic trend, and mention-based user network analysis for presidential issues embedded in Twitter data.

A. Comparison of Term Co-occurrence by Query

Table 1 below compares the partial results of term co-occurrence retrieval by using influential candidates' name, "Geun Hae Park", "Jae In Moon", and "Chul Su Ahn" as query terms.

TABLE I. Term Co-occurrence Comparison By Queries

RK	Query Term and Frequency[1]					
	'Park'	*TF*	*'Moon'*	*TF*	*'Ahn'*	*TF*
1	Presidential Election	634,540	Moon	143,663	**Moon**	**62,770**
2	Declaration of Support	589,728	Presidential Election	92,062	Ahn	53,554
3	Park	530,307	Declaration of Support	62,766	Park	38,750
4	Opinion Poll	492,762	**Ahn**	**57,420**	Presidential Election	35,468
5	Support	82,428	Opinion Poll	54,494	Declaration of Support	11,672
6	Declaration	80,904	Park	49,785	Opinion Poll	10,630
7	Ahn Nonparty	52,990	Democratic United Party	8,552	Democratic United Party	6,858
8	Moon	46,614	Saenuri Party	7,818	Saenuri Party	6,752
9	Ahn	40,078	Moon's Camp	7,796	Moon's Camp	6,540
10	SNS Page	38,748	**Extension of Voting Time**	**7,192**	Moon's Wife	5,740
11	**Park's Mother**	**20,326**	Moon TV	5,916	**Down Contract**	**4,678**
12	**Park's Father**	**13,059**	Moon's Wife	5,768	**Single Candidate**	**3,240**
13	Park's Policy	10,210	**Single Candidate**	**4,140**	Discerning Interview	2,702

[1] The query terms 'Park', 'Moon', and 'Ahn' stand for presidential candidates"Geun Hae Park", "Jae In Moon", "Chul Su Ahn".

Across the board, general terms related to the event of the presidential election such as 'Presidential Election', 'Declaration of Support', 'Opinion Poll', and each candidate's counterparts' names were occurred in common and frequently. On the other hand, unique terms by queries also appeared - they are bolded above. Terms such as 'Park's Father and Mother' from the query 'Guen Hae Park', 'Extension of Voting Time' and 'Single Candidate' from the query 'Jae In Moon', and 'Down Contract' and 'Single Candidate' from the query 'Chul Su Ahn' are the distinct features. In reality, events related to these unique key-words were main issues to each candidate. In particular, what is worth observing carefully is that in tweets which mentioned 'Jae In Moon' and 'Chul Su Ahn', each other's name was occurred with higher frequency and ranking than those on postings which mentioned 'Guen Hae Park'. This is because 'Jae In Moon' and 'Chul Su Ahn' have discussed single opposition candidacy.

B. Social Issue Trend Analysis with DMR

Temporal topic trend on 2012 Korean presidential election shows in Figure 7 and Table II (t = 10, n = 20). Depending on the change of the probability distribution, each topic shows two types of patterns: 1) Rising tendency - topic #1, 2, and 2) Falling tendency - topic #3, 4.

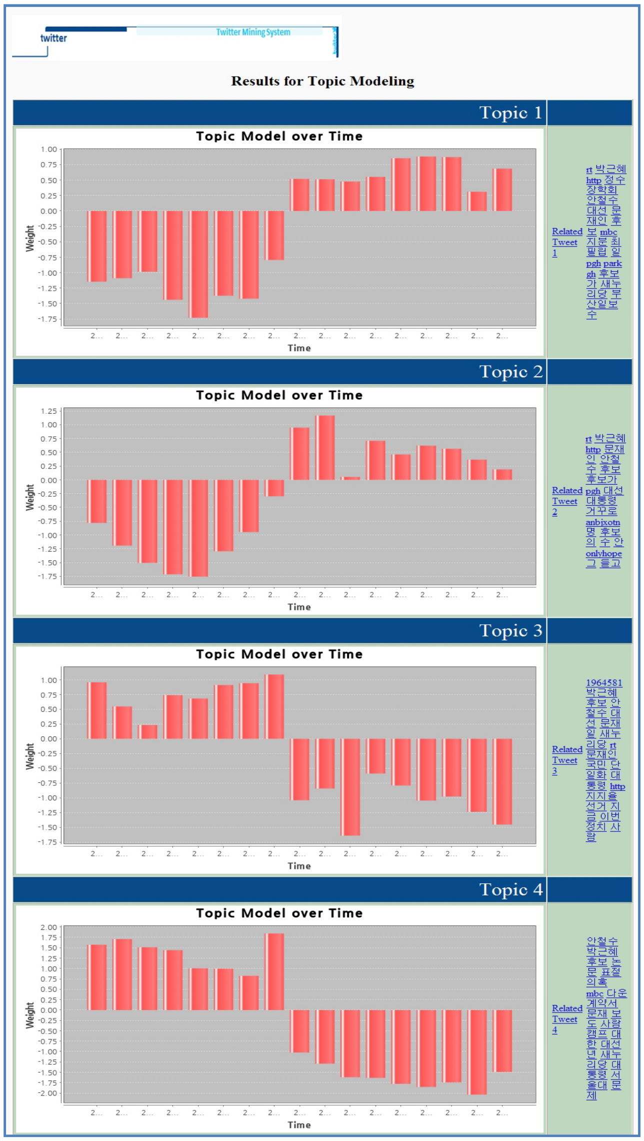


Fig. 7. Topic Trends on the Presidential Election Overview

TABLE II. TOPIC ANALYSIS

Topic	Description		
	Label	***Terms***	***Type*[1)]**
01	Jeongsu Foundation	Park, Jeongsu Foundation, MBC, Pillip Choi	R
02	Candidate	Park, Moon, Ahn, Candidate, Election	R
03	Park's Approval Rating	Park, Single Candidate, Approval Rating	F
04	Suspicion on Ahn	Dissertation, Plagiarism, Down Contract	F

1)' Type' indicates 'R'ising or ' F'alling tendency of a specific issue.

To identify relationship between topical trend in Twitter and changes of social issues in the real world, we compared the topic #1 and #3, with actual events for the same period.

First, related to the topic #1 labeled as 'Jeongsu Foundation', scandals on 'Jeongsu Foundation' and resignation of its board of directors received with media coverage on October 15. At increasing pace, the probability distribution of the topic in Twitter also changes, and even grows faster from October 10. Based on this observation, it indicates that controversial issues are propagated faster in Twitter than other media, and Twitter shows a predictive nature.

Second, during this period, suspicion on 'Chul Su Ahn' (Oct. 1), and Ahn's elucidation for the truth were covered by mass media. This social issue, however, evened out soon because of the explanation of the Ahn's election camp. The topic #3 also shows falling tendency. Based on this, we identified that Twitter would be useful for keeping track of changes of topical influence.

C. Mention-based User Network Analysis

User network in Twitter has different characteristics from those of other social media because users in Twitter can make a relationship by sending-receiving mentions between them. This study, therefore, aims at visualizing and analyzing content-based networks of 136,754 users. Mention-based user network is analyzed by the following process. First, we use a query such as 'Geun Hae Park' to retrieve tweets that contain the query term. Then, we extract the list of users who posted those tweets. Finally, directed user network is visualized, based on the list of users who mentioned them or were mentioned by them.

Network Visualization by Queries 'Guen Hae Park' shows in Figure 8. Based on analysis of user profiles and mentions, it turns out that A and B clusters of Park's network are a set of users who have political leaning toward conservation. The other clusters consist of progressive ones. This result reveals that unlike a certain candidate's political identity, users in Twitter also mention about those who are different from their political dispositions. In addition, conservative-minded users interact with each other, consisting of a few large communities, while the counterparts forming several small groups are connected by some nodes with high betweenness.

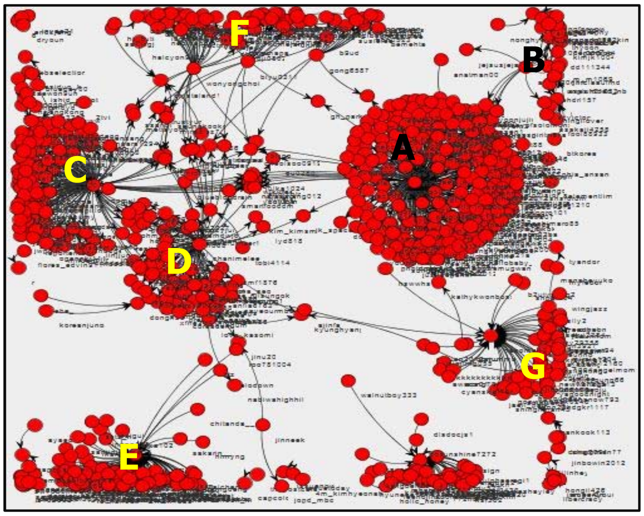


Fig. 8. Network Visualization by Queries 'Park'

In addition, we examine how strongly the groups of the mention-based network are structured. To this end, we use modularity. Networks with high modularity tend to show a dense connection between the nodes within groups where a low modularity indicates sparse connections between nodes in different groups [17]. Modularity calculation for the entire network by JUNG and detected communities by the Voltage-clustering algorithm are shown in Table III.

TABLE III. RESULT OF COMMUNITY DETECTION

No	Community Size	Modularity[1)]
01	33,393	1.25E-04
02	27,666	
03	38,987	
04	32,648	
05	3,760	
06	208	
07	20	
08	15	
09	16	
10	1	
11	1	
12	3	
13	1	
14	35	

1) The value of the modularity lies in the range between -0.5 and 1

As shown above, we generated fourteen distinguished communities by the Voltage-clustering algorithm where the entire network has the value of 1.25E-04 in terms of modularity. This community structure reveals that 1) the user communities on the issue of the presidential election are quite densely connected with a few sparse connections between groups and 2) the network of social interests divides naturally into subgroups. This result also indicates that detecting community structure is a useful approach of as a data analysis technique used to shed light on the structure of large-scale network data sets.

V. Conclusion

In this study, we empirically developed the Real-time Twitter Trend Mining (RT^{2}M) system that is designed for in real-time to 1) crawl and store every tweet produced in Twitter, 2) keep track of topical trend, and 3) visualize mention-based user networks. The major contribution of the study is making it possible to mine dynamic social trends and content-based networks generated in Twitter through adequate integration of state-of-the-art techniques. We also demonstrated the case study on 2012 Korea presidential election. Our empirical study revealed some new findings. First, based on the observation of the temporal Topic Modeling, we identified that Twitter would be a useful medium for keeping track of topical trends. In particular, we found that controversial issues could be propagated faster in Twitter than other media. Second, the mention-based user network provides a basis for identifying any nodes with high betweenness because users in Twitter tend to send/receive mentions each other, not depending on their positive/negative attitude towards a certain issue. This behavioral trait also affects to density of the user community. Through additional sentiment analysis on tweets and sent-received mentions, however, we need to properly measure connectivity among users which could vary based on personal leaning or attitude on a certain issue in Twitter.

In the future, we will study how to apply sentiment analysis to Twitter data to observe changes in public opinion and the formation process of a certain issue, and ultimately design the prediction model of social issues on social media. In addition, by applying the concept of citation to Twitter, we may be able to discover content-based influencers and opinion leaders. We will also explore whether the community structure detected is co-related with the results of topic modeling reported in this paper. In addition, it is one of the interesting research problems to examine the social network patterns varying with the change of different terms over time that we hope to tackle in the future.

Acknowledgement

This study was supported by National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-2012S1A3A2033291).

References

[1] Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 993-1022.

[2] Boyd, D. M., & Ellison, N. B. 2007. Social network sites: definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1). Available at http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html

[3] Cha, M., Haddadi, H., Benevenutoz, F., and Gummadi, K. P. 2010. Measuring user influence in Twitter: the million follower. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.

[4] Chaney, A., Blei, D. 2012. Visualizing topic models. Proceedings of the 6th International AAAI Conference on Weblogs and Social Media.

[5] Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. 2009. Reading tea leaves: how humans interpret topic models. Neural Information Processing Systems.

[6] Choi, D., Min, M., Kim J., and Lee, J. 2011. A Study on topic tracking using microblog. Proceedings of KIIS Spring Conference 2011, 21(1): 80-82.

[7] Drezner, D. W., and Henry F. 2007. Introduction - blogs, politics and power: a special issue of public choice. Public Choice, 134(1-2): 1-13.

[8] Huberman, B. A., Romero, D. M., and Wu, F. 2008. Social networks that matter: Twitter under the microscope. Available at http://ssrn.com/abstract=1313405

[9] Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. 2009. Twitter power: tweets as electronic word of mouth. JASIST, 60: 1-20.

[10] Jansen, H. J., and Koop. R. 2005. Pundits, Ideologues, and Ranters: the british columbia election online. Canadian Journal of Communication, 30(4): 613-632.

[11] Java, A., Song, X., Finin, T., and Tseng, B. 2007. Why we twitter: understanding microblogging usage andcommunities. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and socialnetwork analysis, 56-65.

[12] Kim, Y. 2011. Prediction of structure of spread of public opinion at Twitter with special emphasis on by-election for Seoul mayor. Political communucation, 23: 103-139.

[13] Kwak, H., Lee, C. Park, H., and Moon. S. What is Twitter, a social network or a news media? Proceedings of the 19th International conference on WWW, 591-600.

[14] Lee, K., Namgoong, H., Kim, E. H., Lee, K., and Kim, H. 2010. Analysis of multi-dimensional interaction among SNS users. Journal of Korean Society for Internet Information, 12(2): 113-122.

[15] Livne, A., Simmons, M., Adar, E., and Adamic, L. 2011. The party is overhere: Structure and content in the 2010 election. Proceedings of 5th ICWSM. Available at http://bit.ly/q9lSu

[16] Mimno, D. M. and McCallum, A. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. UAI 2008, 411-418.

[17] Newman, M. E. J. 2006. Modularity and community structure in networks. Proceedings of National Academy of Sciences, 103(23): 8577–8696.

[18] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. 2010. From tweets to polls: linking text sentiment to public opinion timeseries. Proceedings of 4th ICWSM, 122-129.

[19] Hong, L., Davison, B., D. 2010. Empirical study of topic modeling in Twitter. SOMA '10 Proceedings of the First Workshop on Social Media Analytics, 80-88.

[20] Sakaki, T., Okazaki, M., and Matsuo, Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. WWW '10 Proceedings of the 19th WWW, 851-860.

[21] Salton G., McGill M. J. 1986.Introduction to modern information retrieval. NY: McGraw-Hill.

[22] Seol, k., Ki, J., Lee, C., and Baik, D. 2012. Intimacy measurement between adjacent users in social networks. Journal of KIISE, 39(4): 335-341.

[23] Tamer, A. Wang, N., Hale, S., and I Graham, M. 2012. Obama wins the election! (on Twitter). Available at http://www.zerogeography.net/2012/11/obama-wins-election-on-twitter.html

[24] Tumasjan, A. Sprenger, T. O., Sandner, P. G., and Welpe, I. M. 2010. Predicting elections with Twitter: what 140 characters reveal about political sentiment. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, 178-185.

[25] Williams, C., and Gulati, G. 2008. What is a social network worth? Facebook and vote share in the 2008 presidential primaries. In Annual Meeting of the American Political Science Association, 1-17.

[26] Woodly, D. 2007. New competencies in democratic communication: blogs, agenda setting and political participation. Public Choice, 134(1-2): 109-123.

[27] Wu, F., Huberman, B. A. 2004. Finding communities in linear time: a physics approach. The European Physical Journal B - Condensed Matter and Complex Systems, 38(2): 331-338.

[28] Xu, X., Yuruk, N., Feng, Z., Schweiger, T., A., J. 2007. SCAN: a structural clustering algorithm for networks. KDD '07 Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 824-833.

[29] Yan, E., Ding, Y. and Sugimoto, R. 2012. Topics in dynamic research communities: an exploratory study for the field of information retrieval. Journal of Informetrics, 6(1): 140-153.

[30] Yoon, M., Lee, J. 2012. Antecedents of social media use, actual use, and social capital: an analysis of a mediation model. Broadcasting and Communication, 13(2): 5-44.

[31] Zhao, W., X., Jiang, J. et al. 2011. Comparing Twitter and traditional media using topic models. ECIR '11 Proceedings of the 33rd European conference on Advances in Information Retrieval, 338-349.