Analysis of Twitter Research Trends based on SLR

Ilkyu Ha*, Hohwan Park*, Chonggun Kim**

*Department of Computer Engineering, Yeungnam University, Gyeongsan, Gyeongbuk 712-749, Republic of Korea

**Corresponding author, Department of Computer Engineering, Yeungnam University, Republic of Korea

ilkyuha@ynu.ac.kr, hohwan.park@elyasion.com, cgkim@yu.ac.kr

Abstract— **With the increasing interest in micro blogging services, extensive research has focused on Twitter for a variety of purposes. For effective research on Twitter, an investigation of academic research papers related with Twitter will be required in order to analyze research trends. However, it is not easy to review the vast data. Therefore, we extract the representative literature related to Twitter and investigate the trends in Twitter research by using the systematic literature review (SLR) method. The five most popular resource sites are selected to collect Twitter-related literature and the 106 papers selected based on SLR are analyzed to study interesting trends in Twitter research. To classify the research area, the selected papers are divided into five main research categories according to research topics. In addition, other valuable results such as the inclination of the authors are also understood.**

Keywords—**Twitter, Twitter research, tweet, research trend, SNS, SLR**

I. INTRODUCTION

Twitter, a popular micro blogging tool, is strong growing and widely using to expand information since its launch in October, 2006. The number of Twitter users is also steeply increasing. With its variety of applications, many academic researches have focused on Twitter. To understand the trend of such research, searching the relevant literature and analyzing the related papers may be helpful. However, it is not easy to review the vast literature. Therefore, this paper instigates to reveal trends in the recent study of Twitter by the systematic literature review (SLR) method.

The SLR method is adopted as a useful method for concentrating representative papers in this study[1-2]. Five representative Internet search sites that provide research literature are selected. After selecting the appropriate amount of literature, trends in the study of Twitter are analyzed and identified. The selected literature is divided into two groups to investigate intensively the total research trends and latest research trends: groups 1 include studies from 2007 to the present and group2 include studies from 2011 to the present, respectively.

II. RELATED WORKS

SLR [1] has been accepted as a prominent method to classify and extract the valuable literature. According to Kitchenham [1], SLR is a mean of identifying, evaluating and interpreting all available research relevant to a particular research question, topic area, or phenomenon of interest. SLR is composed of three main review phases: planning, conducting, and reporting. The conducting process is expanded and refined up to 4 activities to review the literature[2]. Several studies have used the SLR method to classify study area and find study trends in computer engineering. For example, [2] has used the method to classify studies of UML(unified modeling language) and [19] has used the method to find the trend of UML studies. In the meantime, though a number of studies have studied on Twitter as a hot issue, the study that classify twitter studies and find study trend can not be find. Therefore we start this study to classify Twitter studies using SLR, and then we obtain the recent trend and propose the direction of future study on Twitter. In this study, six procedures based on SLR are modified, as words in parentheses shown in Figure 1.

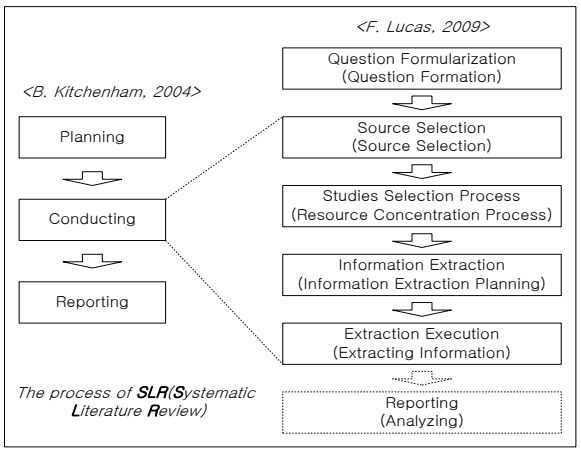


Figure 1. The process of SLR(Systematic Literature Review)

III. SYSTEMATIC LITERATURE REVIEW PROCESS

In this section, we show procedures of SLR which are performed.

A. Question Formation

The following questions are defined in order to attain some analytic results through the SLR process [2].

1. What is being studied about Twitter? This is refined into the following specific questions.

-What are the main topics in Twitter research?

-What are the specific study areas in each research area?

2. What are characteristics of the authors?

-What is their field of study, region, nationality, etc.?

B. Source Selection

The following five renowned literature sources are selected; IEEE Xplore[19], ACM Digital Library[20], Google Scholar[21], Science Direct[22], Web of Science[23].

In the selected sources, the important keywords that can access the highest number of useful results are 'twitter' or 'tweet'. Even though the term 'tweet' is used in the field of zoology as a different meaning, most of the retrieved literature is related to the Twitter service.

C. Resource Concentration Process

The papers retrieved from the five sources are concentrated by criteria. The papers are retrieved twice on two different dates: June 26, 2012 and February 13, 2013. The three resource concentration stages are applied;

(1) First stage: The 'Twitter' or 'tweet' keyword is used to search based on the paper title or abstract from the five sources.

(2) Second stage: By considering the citation counts, some representative papers of each site are selected from the searched papers of stage 1.

(3) Third stage: The same papers from different resources are cross-checked. The overlapped papers that are searched in more than two source sites are included as candidates.

Finally, the final papers that will be examined in this study are selected as shown in the 'final' column of Table 1. 71 papers in group1 and 35 papers in group2 are selected.

TABLE 1. SELECTED PAPERS AT EACH STAGE OF THE SELECTION PROCEDURE

Resource	Group	Stage 1		Stage 2	Overlapped	Stage 3	Final
		2012.6	2013.2				
IEEE	group1	113	196	10	3	7	7
	group2	74	156	4	0	4	4
ACM	group1	358	556	26	24	2	2
	group2	223	358	11	8	3	3
GOOGLE	group1	3960	7040	60	39	58	49
	group2	1680	3210	23	16	23	23
SCIENCE	group1	76	108	8	7	6	5
	goup2	44	76	4	3	2	2
WEB	goup1	262	368	28	15	13	8
	goup2	155	251	14	9	5	3
TOTAL	goup1 (2007~present)	4769	8268	132	88	86	71
	goup2 (2011~present)	2176	4051	56	36	37	35

D. Information extraction planning

Information extraction principles from the selected papers are developed based on the research questions defined in Section Ⅲ.A The following data are extracted: resource name (source site where the paper is found), paper title, authors, citation count of the paper, sort of paper (journal/proceeding), main category, sub-category and characteristics of authors (author's major, nationality, affiliation, announcement year, number of authors).

By classifying the selected papers based on 'study area', five main categories of research purpose are derived. Each main category includes several sub-categories. Table 2 shows 5 main categories and the sub-categories of them.

TABLE 2. MAIN CATEGORIES AND SUB-CATEGORIES

Main category	Sub-category
Trend to use	Adolescent's use(gender difference, potential health outcomes), Imagined Audience(contents producer navigation), Informal Communication(Twitter's role), Information Contagion(spread of news), Information flow(2011 Tunisian revolution), Modeling Dialogue Action(user's action flow), Retweet(why, factors), Trend Detection(identifying emerging topics), Twitter Using(317 users), User Modeling(user activity), Topic(topic comparing)
User relationships	Community Structure, Influencer, Follow relationship, Power user (relationship of elite user), Social network(interaction pattern, student interaction), tie formation(transitivity, mutuality), User influence (measuring influence)
Applications	Commerce(brand sentiment), Company(stakeholder), Crisis communication(reaction), Crowd-sourced Sensing(weather radar), Disaster prevention(earthquake, H1N1, Flu), Education(engagement, grade), Entrepreneur(social interaction), Health care(public health, ALS), News(recommending topic news, obtain breaking news), Politics(political sentiment, congress), Public relations(Haiti), Stock Market, Robotics, Stock Market(predictor)
Twitter Content Analysis	@ sign analysis, Classifier & labeler(recognizing named entities), Event Detection(first story detection), Information Filtering(short text classification), Lexical normalization (short text message), Location Information, Opinion and Sentiment(sentiment analysis, opinion mining), Ranking(recency ranking), Spam detection, Tag Analysis, Text Analysis(part of speech tagger), Social impact(citation)
Effective Use	Cloud computing(data analysis), Information credibility(finding newsworthy topics, credibility analysis), Twitter network analysis (network properties), Technology Management(place topical theme)

E. Extracting information

We carefully read the selected papers in the process of extracting information. To extract information exactly, we prepare for a data sheet for counting and recording the paper information.

IV. ANALYSIS RESULT OF THE SLR

The results obtained from SLR are summarized. In the type of publishing, journal papers and conference proceeding papers are an absolute majority.

A. Study area

Firstly in the study area, 'Application' and 'Twitter content analysis' were the most popular study areas. Figure 2 shows the number of papers based on main topics in each year.

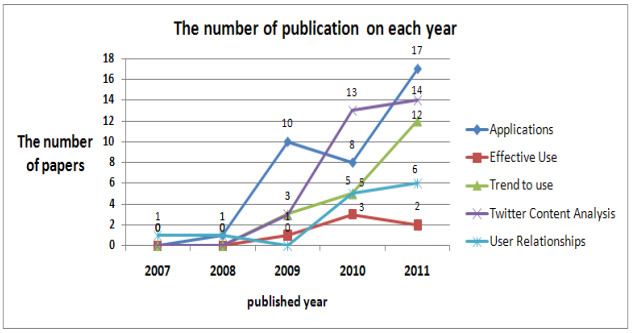


Figure 2. Number of papers in each study area

In the 'Application' main category, the papers related to 'Disaster prevention'[3-5], 'Health care'[6-7] and 'Politics'[8-9] showed a relatively high proportion in group 1 as Figure 3. 'Disaster Prevention' and 'Stock Market' showed a high proportion in group 2. In the 'Trend to Use' main category, the papers related to 'Retweet'[10], 'Twitter Using'[11] and 'User Modeling'[12] had the higher proportion. In 'Twitter Content Analysis' main category, the papers related to 'Opinion and Sentiment'[13-14], 'Spam detection'[15] and 'Tag analysis'[16] showed a higher proportion, as shown in Figure 4.

As mentioned in section 3.3, papers are retrieved from the source sites on two different dates: June 26, 2012 and February 13, 2013. Between these two dates, the number of papers has increased rapidly. Especially, 'Applications (group 2)' and 'Twitter content analysis (group 2)' show strong growth in citation count, as shown in Figure 5.

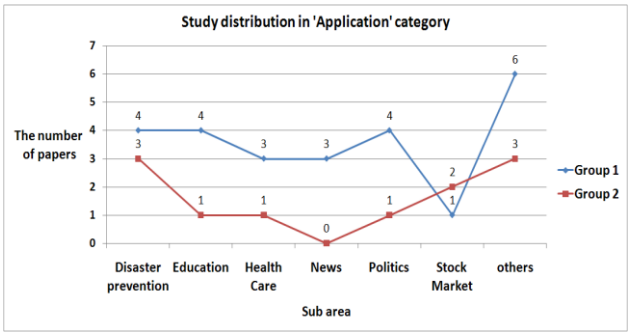


Figure 3. Study distribution in 'Application' category

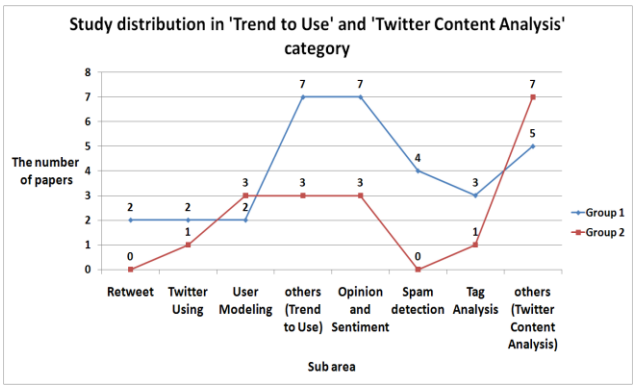


Figure 4. Study distribution in 'Trend to Use' and 'Twitter Content Analysis' categories

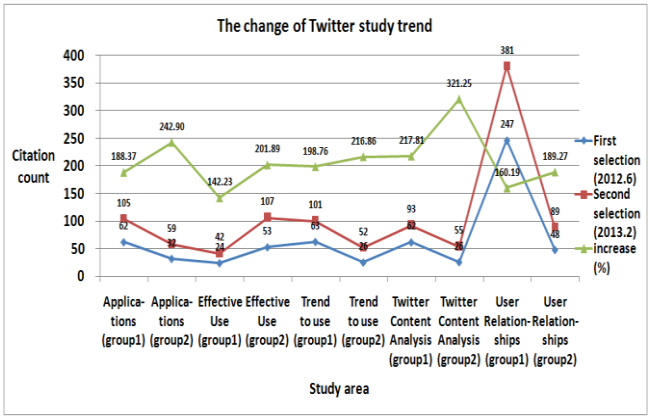


Figure 5. Twitter research trends

B. *The scale and sampling period of tweets*

The scale of tweets such as the number of users, the number of sampling tweets and the sampling period on each paper was outlined. 'Tag analysis', 'Spam detection' and 'Disaster prevention' studies of group 1 examined more user accounts as shown in Figure 6. 'Tag analysis' and 'Spam detection' studies of group 1 treated more tweets as shown in Figures 7. The average of the tweet sampling period is about 120 days. The longer tweet collection period is shown in the 'Spam detection' and 'Tag analysis' study areas in group 1 as shown in Figure 8. In the case of group 2, the result is similar to the tendency of group 1.

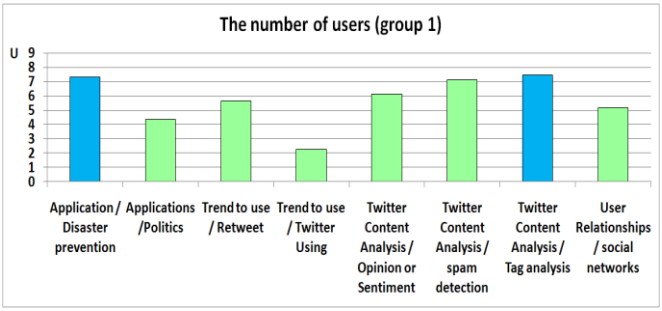


Figure 6. The number of user accounts (group 1) ($U=\log 10^{x}$)

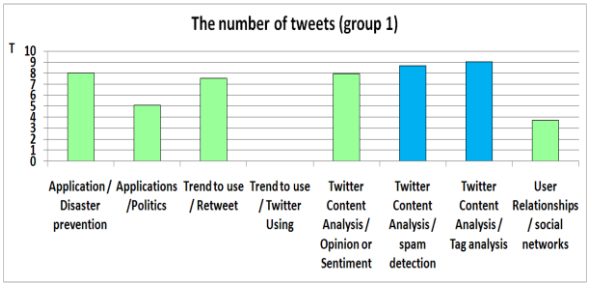


Figure 7. The number of tweets (group 1) ($T=\log 10^{x}$)

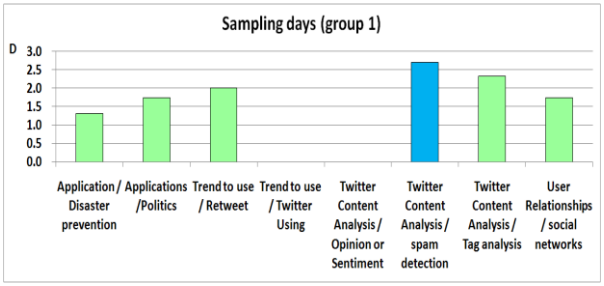


Figure 8. Sampling period (group 1) ($D=\log 10^{x}$)

C. *Author's information*

The author information of the papers such as major, nationality, affiliation, and the published year was analyzed. Most of working fields of authors are related to computer science or IT (Information Technology) as shown in figure 9. In the case of group 2, the result is similar to that of group 1. The most active country on Twitter study is U.S. and the number of published papers is increasing rapidly.

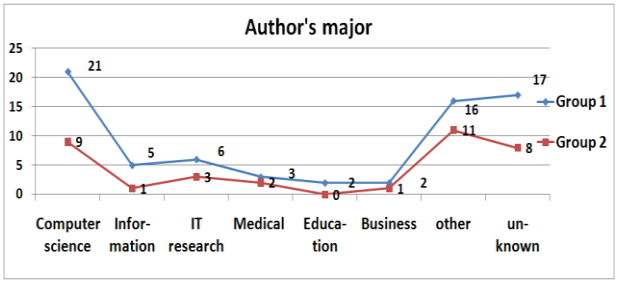


Figure 9. Author's major

D. *Trends on Twitter research*

The results of this study show the following trends.

(1) Most of the studies are focused on 'Application', 'Trend to use' and 'Twitter content analysis'. These studies are analyzing the flow and content of tweet rather than any technical problems on Twitter.

(2) Twitter has been applying in a variety of fields including education, politics, economy, health care and natural disaster prevention[17-18]. Such applications required a lot of data for getting valuable information.

(3) The studies on the effective usage of Twitter such as 'Ranking', 'Spam detection', and 'Event detection' based on the content of tweets are very valuable. Despite the extensive efforts are necessary to gather tweet data and user accounts, the studies related to the category are increasing.

(4) The studies that extract user's relationship from the flow and content of tweet such as 'Influencer', 'Power user', and 'Follow relationship' are also important. This is a unique study area that must be studied on Twitter study.

(5) Most common major field of the authors is related to IT. The majority of the authors belonged to U.S. institutions. The number of papers about Twitter has been continuously increasing since 2007.

V. CONCLUSIONS

The study purpose was to examine and evaluate the current trends of Twitter research by using the SLR method. A total of 12,319 published papers were extracted from the five most recommendable public resource sites on two different dates: June 26, 2012 and February 13, 2013. By the proposed SLR method, 106 papers were finally selected for detail analysis. By analyzing the selected papers, the trends of Twitter research and characteristics of author were identified.

Most of the authors express positive opinions regard to the present and future of Twitter. They expect continuous improvement in Twitter itself and its application in a variety of fields. The future of Twitter research may be expanded to a variety of study areas based on the present trends.

ACKNOWLEDGMENT

This work has been funded by the BK21+ program of the National Research Foundation (NRF) of Korea.

REFERENCES

[1] B. Kitchenham, "Procedures for Performing Systematic Reviews", NICTA Technical Report, 2004.

[2] F. Lucas, F. Molina, A. Toval, "A systematic review of UML model consistency management", Information and Software Technology, Vol.51, pp.1631-1645, 2009.

[3] T. Sakaki, M. Okazaki, Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors", Proc. of WWW 2010, pp.851-860, 2010

[4] A. Hughes, L. Palen, "Twitter adoption and use in mass convergence and emergency events", International Jour. Of Emergency Management, Vol.6, No.3-4, pp.248-260, 2009

[5] C. Chew, G. Eysenbach, "Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak", Research Article of PLOS ONE, Vol.5, No.11, 2010

[6] M. Paul, M. Dredze, "You are what you tweet: Analyzing Twitter for public health", Proc. of the Fifth International AAAI Conference on Weblogs and Social Media, pp.265-272, 2011

[7] C. Hawn, "Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care", Health Affairs, Vol.28, No.2, pp.361-368, 2009

[8] A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment", Proc. Of the Fourth International AAAI Conference on Weblogs and Social Media, pp.178-185, 2010

[9] J. Golbeck, J. Grimes, A. Rogers, "Twitter use by the US Congress", Jour. Of the American Society for Information Science and Technology, Vol.61, Iss.8, pp1612-1621, 2010

[10] D. Boyd, S. Golder, Gilad Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter", Proc. Of HICSS 2010, pp.1-10, 2010

[11] W. Reinhardt, M. Ebner, G. Beham, C. Costa, "How people are using Twitter during conferences", Proc. of 5th EduMedia conference, pp.145-156, 2009

[12] K. Chretien, J. Azar, T. Kind, "Physicians on Twitter", Jour. of JAMA, Vol.305, No.6, pp.566-568, 2011

[13] A. Pak, P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining", Proc. of LREC 2010, 2010

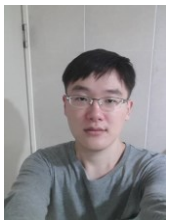
[14] J. Bollen, A. Pepe, H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena", Proc. of the Fifth International AAAI Conference on Weblogs and Social Media, pp.450-453, 2011

[15] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, "Detecting spammers on twitter", Proc. of CEAS 2010, 2010

[16] D. Romero, B. Meeder, J. Kleinberg, "Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter", Proc. of WWW'11, pp.695-704, 2011

[17] A. Vivacqua, M. Borges, "Taking advantage of collective knowledge in emergency response systems", Jour. of Network and Computer Applications, Vol.35, Iss.1, pp.189-198, 2012
[18] M. Domingo, "An overview of the Internet of Things for people with disabilities", Jour. of Network and Computer Applications, Vol.35, Iss.2, pp.584-596, 2012
[19] IEEE Xplore: http://ieeexplore.ieee.org/
[20] ACM Digital Library: http:// http://dl.acm.org/
[21] Google Scholar: http://scholar.google.co.kr/
[22] Science Direct: http://www.sciencedirect.com/
[23] Web of Science: http://apps.webofknowledge.com/

Ilkyu Ha Received the PhD degree in computer engineering from Yeungnam University, Korea. He is currently a visiting professor in computer engineering department at Yeungnam University. He was a software developer at HSS of Korea, a senior researcher at center for innovation of engineering education at Yeungnam University. His research interests include Information Networks, Internet Applications and software engineering



Hohwan Park He is a master degree course student in computer engineering department at Yeungnam University. His research interests include Information Networks, Sensor Network and social networking services.

Chonggun Kim Received the PhD degree in computer science from the University of Electro-Communications, Japan. He is a professor in Computer Engineering department at Yeungnam University. He was a research fellow at UCSC and Virginia Tech, USA. He is currently interested in Information Networks, Internet Applications and social networking services. He is a member of the IEEE and the IEEE Computer Society.