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
Tweet, a popular micro blogging tool, is strong growing and widely used to spread 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 group 2 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.

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?
   - 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 group 1 and 35 papers in group 2 are selected.

| TABLE 1. SELECTED PAPERS AT EACH STAGE OF THE SELECTION PROCEDURE |
|-------------------|-------------------|-------------------|-------------------|-------------------|
| **Resource**      | **Group**         | **Stage 1**       | **Stage 2**       | **Overlap**       |
|                   |                   | **2012**          | **2013**          | **Stage 3**       |
|                   |                   |                   |                   | **Stage 4**       | **Final**        |
| 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               |
|                   | group2            | 44                | 76                | 4                | 3               | 2               | 2               |
| WEB               | group1            | 262               | 368               | 28               | 15              | 13              | 8               |
|                   | group2            | 155               | 251               | 14               | 9               | 5               | 3               |
| TOTAL             | group1            | 4769              | 8268              | 132              | 88              | 86              | 71              |
|                   | group2            | 2176              | 4051              | 56               | 36              | 37              | 35              |
| (2007–present)    |                   |                   |                   |                   |                 |                 |                 |
| (2011–present)    |                   |                   |                   |                   |                 |                 |                 |

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>
<thead>
<tr>
<th>TABLE 2. MAIN CATEGORIES AND SUB-CATEGORIES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main category</strong></td>
</tr>
<tr>
<td><strong>Adolescent’s use</strong></td>
</tr>
<tr>
<td><strong>Trend to use</strong></td>
</tr>
<tr>
<td><strong>User relationships</strong></td>
</tr>
<tr>
<td><strong>Applications</strong></td>
</tr>
<tr>
<td><strong>Effective Use</strong></td>
</tr>
</tbody>
</table>

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.
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.

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.
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.

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


[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.