Vol.2014-DBS-159 No.13
Vol.2014-IFAT-115 No.13
2014/8/2

Topics and Influential User Identification in Twitter using Twitter Lists

Guanying Zhou[†1] Hiroki Asai[†1]
Hayato Yamana[†1,2]

Twitter, as one of the most popular social network services, draws the attention of more and more researchers worldwide. With a large amount of information tweeted every day, it turns essential to identify the influential users we are interested in. In the previous research, researchers mainly identify topics from tweets and rank users by utilizing the follow relationship; however, the following relationship is strongly related to their reputation in real world and cannot describe their influence and activity level in Twitter exactly. Instead, in this paper, to identify topics and influential users, we use "Twitter List," whose name represents the topic of listed members. By analyzing Twitter List, we are able to detect topics and identify influential users in the corresponding topic more efficiently. Based on our experimental evaluation using the selected two topics, the influential users identified by our proposed method have the average influence score related to the topic made by interviewees of 3.7 and 3.33 outweigh the methods of ranking by follower numbers with the average score of 3.22 and 3.27 respectively.

1. Introduction

Twitter is one of the most popular social network services with many registered users from all over the world. In Twitter, registered users can post 140-character messages, follow other users, retweet tweets published by other users, reply to other users, comment tweets and enjoy many other social activities in the platform of Twitter.

The platform of Twitter was created in the year of 2006, and gained popularity all over the world rapidly. According some technical reports[ab], the platform of Twitter have 500 million registered users in 2012, 340 million tweets posted per day. The service also handled 1.6 billion search queries per day[c].

In the year of 2009, Twitter started the new function of "Twitter List", which allows users to access the tweets of a group of users only by following the list. A Twitter List is a curated group of Twitter users. In order to create a Twitter List, first of all, the creator should name the List, then include other users into the List. The name of the Twitter List describes the common attribute of the users in the List. From the timeline of the List, users can view the tweets published by all of the members in the List just from the timeline of the List. Users can also subscribe Twitter List created by other users.

Users utilize the function of Twitter List for the following reasons[d]: 1) Monitor without following. Users can monitor a group of individuals in a list without following any of the folks on the list. 2) Easy user management. Users can find a list and follow the entire listed users with one button, which makes it easy to manage. 3) Promote lists. It is a great way for a group of users to promote each other in Twitter by creating one list. 4) Build a bigger following circle. Creating a handful of useful list, promoting the lists in public places and the member on the list will raise your exposure and lead to more followers.

According to the survey of the research firm Pear Analytics[e], only 9% of the tweets in Twitter have values that should be paid attention to others are useless information. So it is essential to obtain information with value from Twitter in an efficient way. One of the methodologies to solve the problem is to follow the users who are authorities and often tweet valuable information in the area. The motivations of our research are to extract topics and identify influential users in Twitter.

In the previous research, researchers mainly extract the interest topics from the tweets published by users and rank users utilizing the follower/friends graph or retweet graph. In Twitter, some of the users rarely publish messages with value but got many followers because of their influence in real life. So the follower/friend graph cannot represent their real influence in Twitter. In our research, we utilize the Twitter List, which can better represent their influence in Twitter, to extract topics and rank users.

The rest of the paper is organized as followings: Section 2 introduces some related research in recent years. Section 3 shows our proposed methodologies. Section 4 shows the results of our experiment followed by our conclusions in Section 5.

2. Related Research

In the previous research, most of the research about topics extraction in Twitter is based on the tweets published by users. And research about influential user identification is based on friend/follower graph or retweet graph.

2.1 Research related to Topics Extraction

Weng et al. [1] extracted user interest topic by collecting the tweets published by the users and applying the LDA model to calculate the probability of user interest.

2.2 Research related to Influential User Identification

Weng et al. [1] proposed an algorithm called TwitterRank to

†1 Waseda University.
†2 National Institute of Informatics

a "Twitter Passed 500M Users in June 2012, 140M of Them in US; Jakarta 'Biggest Tweeting' City". TechCrunch. July 30, 2012.

b Twitter turns six. Twitter.com, March 21, 2012. Retrieved December 18, 2012.

c Twitter Search Team (May 31, 2011). "The Engineering behind Twitter's New Search Experience". Twitter Engineering Blog. Twitter. Retrieved June 10, 2011.

d 5 Reasons to Use the New Twitter List Feature. http://www.ducttapemarketing.com/blog/2009/11/09/5-reasons-to-use-new-the-twitter-list-feature.

e Ryan Kelly, ed. (August 12, 2009). "Twitter Study – August 2009" (PDF). Twitter Study Reveals Interesting Results About Usage. San Antonio, Texas: Pear Analytics. Archived from the original on 2011-07-15.

measure the influence of users in Twitter. TwitterRank measures the user influence taking both the topical similarity between users and the link structure into account. It constructed the graph based on follower or friend relationship.

However, it has been proved that in-degree, i.e. the number of followers, represents a user's popularity, but is not related to other important notions of influence such as engaging audience, i.e., retweets and mentions [2].

Yamaguchi et al. [3] showed the concepts that to identify authoritative users, it is important to consider actual information flow in Twitter and existing approaches only deal with follower or friend relationship among users. They also proposed TURank (Twitter User Rank), in which an algorithm for evaluating users' authority scores in Twitter based on link analysis based on graph called user-tweet graph is introduced.

Kwak et al. [4] ranked Twitter users by the total number of followers, PageRank in the following/follower network and number of retweets in the diffusion network respectively. The results indicate that there exists a gap between the number of followers and the popularity of one's tweets.

Nakajima et al. [5] proposed methods to identify influential bloggers based on the knowledge level of the blogger. It calculates the scores of each blogger in each topic based on the domain-specific words.

Wu et al. [6] ranked Twitter users by the frequency of being listed in each topic category.

2.3 Research related to Twitter List

Yamaguchi et al. [7] proposed a method of tagging Twitter users using Twitter list. The proposed method extracts tags from list names, and tag list members using the tags by analyzing the correlations among twitter lists, tags and users. However, there are many synonymous tags and name tags. Unfortunately, the proposed method ranks a lot of these tags in higher ranks.

Ghosh et al. [8] proposed the method of expert sampling; considering a Twitter user as a 'topical expert' if and only if the user has been listed at least 10 times on some particular topic. The conclusions show that the experts' tweets are significantly richer in information content, cover more diverse topics, and more popular content. Experts' tweets are also more trustworthy and they often capture breaking news stories marginally earlier than random sampling. These properties of expert sampling make it a valuable methodology for generating content for several important content-centric applications, such as topical search, trustworthy content recommendation, breaking news detection, and so on.

Rakesh [10] proposed a method for recommending Twitter List to users considering the personal interests of users. They developed a ListRec model and the network-based LIST-PAGERANK model to recommend Lists to users who are interested in. Our research focuses on ranking Twitter users and recommending them to utilize the Twitter List.

The researches above show that the name of Twitter List is closely related to the interest topic of user and the number of times user has been listed can show the influence of the user. So we begin our research about topics and influential user identification using Twitter lists.

3. Methodologies

3.1 Twitter List

Twitter list, a new function adopted officially by Twitter can help user manage their account. For example, Twitter user can create a list and include other users in the list then they can read the tweets published by the members included in the list directly. Users can also follow lists created by other users to read the tweets.

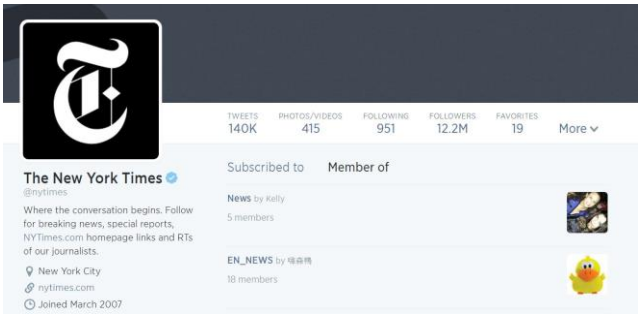

Figure1. The Profile of New York Times

For example, in Figure 1, from the profile page of Twitter user New York Times, we can see many users include the user of New York Times as the list of news.

In the utilization of Twitter List, user can subscribe lists created by other users to read the whole time line tweets of users included in the list. The subscriber count of a List can reflect the influence of the list.

3.2 Dataset

Here we introduce a simple example to explain the words of user, list and tag in our research.

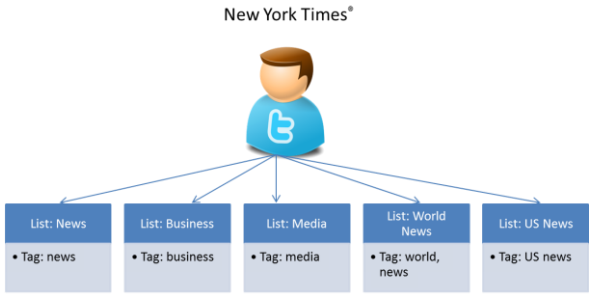

Figure2. Example of Relationship between Twitter user, list and tag

Figure2 explains the relationship between Twitter user, list and tag. For example, the Twitter user of "New York Times", has been listed for 5 times. The name of Lists are "news", "business", "media", "world news", "US news" so we have the tags of "news" for three times, "business" for once, "media" for once, "world" for once and US for once.

In our research, we only focus on the users whose language is English, and the name of the list is English. In our dataset, we collected users whose language is English randomly by using Twitter API. Then, we collected all of the Twitter List in which

the users had been included into. The Summary of collected data is shown below:

- 8,351 users
- 194,369 lists
- 231,376 tags

In the dataset, we only collected users who have been listed at least once. In one list, the name of the list contains several words and we treat one word as one tag. Moreover, in the process of extracting tags from the name of the list, we exclude characters other than English. We have 231,376 tags in total and they contain 37,125 unique English words.

Table1. Top Tags in the Dataset Listed by Frequency

Tag	Frequency
my	8,053
news	5,960
music	4,515
you	2,664
gaming	2,242
i	2,044
the	2,024

In Table1, we list the top tags extracted from the name of all of the lists in our dataset. From the table, we can understand the keyword “news”, “music”, and “gaming” stand for popular topics often appear in Twitter. The keywords “my”, “you”, “i”, and “the” are meaningless stop words, which need to be removed before applying the LDA model.

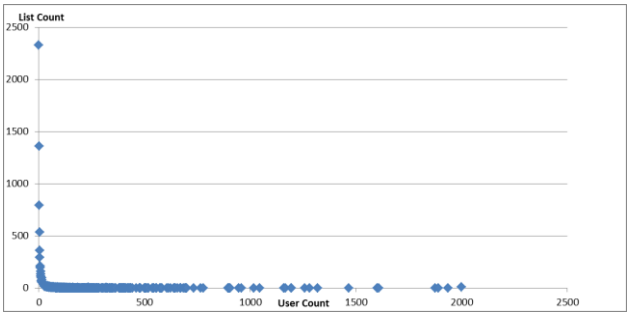

Figure3. The Distribution of Twitter List Count

Figure3 shows the distribution of Twitter List counts of the users in our experiment dataset. From Figure2, we can understand that most of Twitter users have been listed below 500 times and only a few users have been listed for over 2,000 times.

3.3 Methodologies

Our research relies on the official function of Twitter List to identify topic and influential user in Twitter, which has not been done in the previous research. In our previous research [11], we tried to extract possible interest topics from the tweets published by users and rank users based on retweet relationship graph using LDA and PageRank algorithms

Figure4 shows the general flow of our methodologies. Firstly, we obtain the Twitter List resources from Twitter API. Secondly, we extract English tags from the name of the List resources. Next, we apply of LDA model to extract topics from the name of the lists. In the next step of influential user identification, we calculate the topic influence which in related to the subscriber count of Twitter List, and then calculate the user influence. We will introduce each step in details in the following parts.

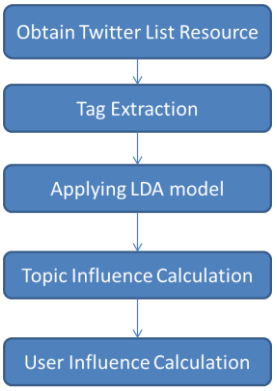

Figure4. General Flow of Methodologies

3.3.1 Obtain Twitter List Resource

We obtain the Twitter List Resource by utilizing the Twitter API[f]. In our research, we mainly use the List API of REST API. We utilize the “GET list/memberships” API to obtain the lists the specified user has been added to. In the process, the id or screen name of the specified user is essential. In the List Resource, the name, slug, member count, subscriber count, information about the creator of the list and other information are provided. In the acquisition of List Resource, we have the important problems of the limitation of Twitter API. In the utilization of “GET list/member ships” API, we can only access it 15 times in every 15 minutes.

3.3.2 Tag Extraction

The creator of the Twitter List names the list and adds other users into the list. The name of the List can reflect the attributes of the members included in the List from some perspectives. We treat every word included in the name of the List as one tag.

3.3.3 Applying LDA model

Latent Dirichlet Allocation (LDA) model [9] is a topic model proposed by David Blei, Andrew Ng, and Michael Jordan in the year of 2003. In the process of applying LDA model, all we need is the set of documents and the number of topics training with. In the results of LDA model, each document is viewed in the form of probability distributions and each topic is assumed to be characterized by the keywords and corresponding probability.

LDA model is widely used for topic extraction from texts. Recently, it is popularly used in the research of Twitter and other social network services for topic extraction.

In our research, we only focus on English words so before applying the LDA model, we exclude the words other than English. Moreover, in order to improve the performance of topics extraction of LDA model, we also exclude the stop words that are meaningless in the process of topic extraction.

In applying the LDA model, the tags extracted from the name of the lists including the same user is regarded as one document and train the LDA model with 50 topics. As the results of the LDA model, we obtain two probability distribution metrics:

f Twitter Developers. https://dev.twitter.com/

$UserTopicProbability_{ij}$ and $KeywordTopicProbability_{kj}$. $UserTopicProbability_{ij}$ represents the probability $User_i$ has interests in $Topic_j$. In the results of the LDA model, the topics extracted from the tags of the list are represents by several keywords and corresponding probability. $KeywordTopicProbability_{kj}$ stands for the probability $Keyword_k$ occupies in $Topic_j$. In the following step, we will utilize these probability metrics to rank Twitter users.

3.3.4 Topic Influence Calculation

In order to evaluate the influence of each topic extracted from the LDA model, we utilize $KeywordTopicProbability_{kj}$ to calculate the influence of each topic.

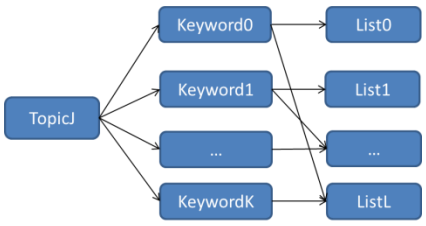

Figure5. Topic Influence Calculation

Figure5 shows the way how we calculate the influence score of each topic. In the influence calculation of $Topic_J$, $Topic_J$ is consist of $k+1$ keywords from $Keyword_0$ to $Keyword_K$ and their corresponding probability $KeywordTopicProbability_{kj}$. In each keyword, we can each keyword to several lists whose name contains the keyword.

$$Influence(Topic_J) = \sum_{k=0}^{K} KeywordTopicProbability_{kj} * \sum_{l=0}^{L} subscriber\_count(List_l)$$

(Equation 1)

Equation1 shows the calculation of topic influence. $subscriber\_count(List_l)$ represents the number of users who subscribe the list which shows the importance of the list.

3.3.5 User Influence Calculation

In order to calculate the influence score of each user and identify the influential users, we utilize $UserTopicProbability_{ij}$ and influence score of $Topic_j$ to calculate the influence score of each user.

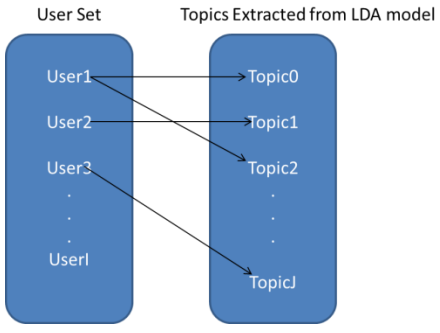

Figure6. User Influence Calculation

Figure6 shows the methodologies of how to calculate the influence score of each user. In the results of the LDA model, $UserTopicProbability_{ij}$ represents the probability $User_i$ has in $Topic_j$.

$$Influence(User_i) = \sum_{j=0}^{J} UserTopicProbability_{ij} * Influence(Topic_j)$$

(Equation 2)

Equation2 shows how to calculate the influence score of each user. We calculate the user general score according to how much probability user has in each topic and the influence score of each topic. In this way, we can rank Twitter users and identify influential users.

4. Results

In this part, we list some of the results of our experiment. In 4.1, we list some of the topics and the corresponding keywords in the results of the LDA model. In 4.2, we list top 10 Twitter users ranked by our proposed methodologies.

4.1 Results of LDA model

Here, we list some of the results of the LDA model. In the results of the LDA model, we train the whole texts formed by the tags of Twitter List with 50 topics. In the results of the LDA model, we can obtain the keywords of each topic and the probability it occupies. Here we list 6 of the 50 topics and the corresponding keywords. We try to analyze the results of the LDA model from the keywords of each topic. Table1 shows the results of the LDA model.

Table2. Topics in the LDA model (Sample)

Topics	Keywords
0	writers, gaming, authors, bloggers, asmsg, writing, political , startup, libertarian, fabulous
1	news, media, general, kesehatan, detik, arab, portal, sehat, detikhealth , magazines
2	health, big, digital, medical, woah, belovedcome, kesehatan, funnys, fitness, godspeedgreatcommission24
3	xbox, battlefield, apps, basketball, garena, notable, georgia, randomness, visual, shopping
4	info, favorite, entertainment, life, public, somerset, local, kesehatan, ios, expert
5	music, dj, berita, funny, art, follow, frasil, ceo, jokes, online

In Table2, we can understand the topics from the related keywords. In Topic 0, the keywords “writers”, “authors”, “bloggers”, and “writing” mean the topic stands for “writing”. In Topic 1, the keywords “news”, “media”, and “magazines” mean the topic stands for “media”. In topic 2, the keywords “health”, “medical”, and “fitness” mean the topic stands for “health care”. In Topic 3, the keywords “xbox”, “apps”, “basketball”, and “visual” mean the topic stands for “video games”. In topic 5, the keywords “music” and “dj” mean the topic stands for “music”.

However, in some of the topics, we cannot understand the topics by analyzing the keywords of the topic, such as Topic 4 listed in Table1. Even though we utilize the name of the Twitter

List, which better describes the attributes of users, not all of the topics are understandable and how to extract more understandable topics remains for future work.

4.2 Top Users ranked by our methodologies

We ranked Twitter users of our dataset by our proposed methodologies described in Section 3.

Table3. Top 10 Users Ranked by Proposed Methodology

Rank	User
1	lawlrah
2	deejieying
3	ebr_ebr
4	NinjaNaii
5	La__Kush
6	AraaK_
7	mouchakisa2
8	ambaaa_007
9	Sweet_Lady_Mara
10	ASVPxSkinny

All of the Twitter users listed above are normal Twitter users, not celebrities or news media. However, they have got a large number of friends and publish many tweets so they take active part in the platform of Twitter. By following these top users who got a higher score over all of the topics, we are more likely to start a conversation and make friends with them.

4.3 Influential Users of Selected Topic

Table4. Influential Users in Each Topic

User Rank	Topic 0	Topic 1
1	Custard_Corner	MalaysiaGazette
2	Deathstrike2014	maalaimalar
3	fitrian	abarandun
4	Braahna	dinamalarweb
5	GhostFaceKeylow	YarmoukNews

In order to verify our proposed method can identify influential users in the understandable topics, we selected the top 5 users in $Topic_0$ and $Topic_1$ of the above table representing "writing" and "news media" topics respectively. Table4 shows the users who have got the top 5 scores in the selected topic.

5. Evaluations

In the previous researches, they often refer influential Twitter users as the users who have many followers. However, the users with many followers are often famous celebrities or news media or organizations. Some of them are inactive Twitter users and they rarely have activities in Twitter. We evaluate our proposed methodologies from two perspectives: 1) Top users with higher scores over all of the topics are active users using the platform of Twitter actively. 2) In each of the selected topic, the top users always publish messages with value related to the topic but get relatively few followers.

In order to verify our proposed method can identify influential users which cannot be identified by ranking by the number of followers, we regard the method of ranking users by the number of followers as baseline and compare the two methods.

5.1 Evaluation by Questionnaire

As tentative evaluation, we selected only the two topics for evaluation. In each topic, we select top 5 users in both our proposed method and baseline. In each of the selected topic, we treat both the top 5 users ranked by our method and the baseline as one group. We asked 6 students to grade each group of users making a score from 1, 2, 3, 4, to 5 according to their influence in the topic, then calculate the average score of each group of users. Here, higher score represents higher influence.

Table5. Questionnaire Results of Each Topic

Topic	0	1
Proposed Method	3.7	3.33
Baseline	3.22	3.27

In each of the selected topic, we calculate the average of each user by the total score received from the respondent divided by the number of respondents. Table5 shows that the top influential users identified by our method have more average scores than the baseline. Our proposed method identified the influential users who often publish related messages but have relatively less followers.

5.2 Evaluation by User Message Amount

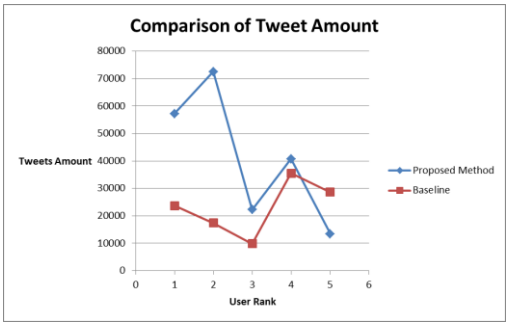

Figure7. Comparison of Tweet Amount

Figure7 shows the comparison of the number of tweets tweeted by the top users ranked by our proposed method and the baseline. From the figure, we can easily understand users ranked top by our proposed method outweigh the baseline from the aspect of published tweet amount. Our method can identify the influential users who take active part interaction in Twitter and publish a large amount of information.

5.3 Evaluation by Topical Difference

In order to prove the specialty of the top users ranked by our proposed method, i.e., to prove that they are expert in a few areas, not those who publish messages randomly in a wide range of areas, we use the topical difference defined in paper [1] to help us evaluate our proposed methodologies and baseline.

Firstly, we normalize the metrics of UserTopicProbability as $UserTopicProbability'$ such that $||UserTopicProbability'_i|| = 1$ for each row $UserTopicProbability'_i$.

Definition: Topic difference between two Twitter user u_i and u_j can be calculated as:

$$\text{dist}(i,j) = \sqrt{2 * D_{JS}(i,j)}$$

(Equation 3)

$D_{JS}(i,j)$ is the Jensen-Shannon Divergence between the two probability distributions $UserTopicProbability'_{i.}$ and $UserTopicProbability'_{j.}$, which is defined as:

$$D_{JS}(i,j) = \frac{1}{2}\Big(D_{KL}(UserTopicProbability'_{i.}||M) + D_{KL}\big(UserTopicProbabilty'_{j.}||M\big)\Big)$$

(Equation 4)

M is the average of the two probability distributions, i.e. $M = \frac{1}{2}\big(UserTopicProbability'_{i.} + UserTopicProbability'_{j.}\big)$. D_{KL} is the Kullback-Leibler Divergence which defines the divergence from distribution Q to P as: $D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$.

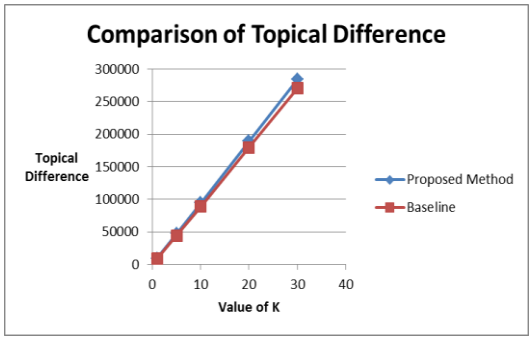

Figure9. Comparison of Topical Difference

We use the methods described above to calculate the topical difference of every pair of users in our experiment dataset. We calculate the total topical difference of top-k users between every other user in the dataset. We select top-k users ranked by both our proposed methodology and baseline.

In Figure9, we can understand the top users ranked our proposed methodologies have more topical difference between other users in the dataset than the top users ranked by the number of followers. It means the top users identified by our proposed methodologies have relatively limit range of topical areas than the users ranked by the number of followers.

We compare the top users with the baseline from the above three perspectives. It has been shown that the users ranked by our proposed methodologies are more active in interaction, more involved in the platform of Twitter and have more topical difference with other users in the dataset.

6. Conclusions

In our research, we proposed a new method of topics and influential user identification in Twitter by using Twitter Lists. Twitter List, which is relatively new function of Twitter, has been studied recently. With the help of the name of list, we can successfully identify the interest topics of Twitter users. Moreover, with the help of subscriber count of Twitter List, namely, the follower of the Twitter List, which reflects the popular level of Twitter List, we identify influential Twitter users who cannot be identified by other methodologies.

From the analysis of the evaluation of our results in Section 5, our proposed methodologies can identify influential user more active in interaction, more involved and show more topical interest compared to the users ranked top by the number of followers. As a conclusion, which is also proved by other papers, there exists a gap between the users who play an important role in Twitter and those users with many followers. In Twitter, many users follow other celebrities or organizations because of their huge influence in real life, not because of their influence of information spread in Twitter.

Our research can identify the users who are really influential in Twitter, not those users with many followers. Especially, we utilize the official function of Twitter List, to help identify topics and influential users, which have not been done in the research before.

Reference

1) J. Weng, E. Lim, J. Jiang, and Q. He , “TwitterRank: Finding Topic-sensitive Influential Twittereres,” Proc. of 3rd ACM Int’l Conf. on Web Search and Data Mining, pp.261-270, New York, USA, February 4-6, 2010.

2) M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, “Measuring User Influence in Twitter: The Million Follower Fallacy,” Proc. of 4th Int’l AAAI Conf. on Weblogs and Social Media, pp.10-17, Washington, DC, USA, May 23-26, 2010.

3) Y. Yamaguchi, T. Takahashi, T. Amagasa, and H. Kitagawa, “TURank: Twitter User Ranking Based on User-Tweet Graph Analysis,” The 11th Int’l Conf. on Web Information Systems Engineering, LNCS, No.6488, pp.240-253, 2010.

4) H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a Social Network or a News Media?” Proc. of WWW 2010, pp.591-600, Raleigh, North Carolina, USA, April 26-30, 2010.

5) S. Nakajima, J. Zhang, Y. Inagaki, T. Kusano, and R. Nakamoto, “Blog Ranking Based on Bloggers’ Knowledge Level for Providing Credible Information,” Proc. of the 10th Int’l Conf. on Web Information Systems Engineering, LNCS, No.5802, , pp.227-234, 2009.

6) S. Wu, J. M. Hofman, W. A. Mason, D. J. Watts, “Who Says What to Whom on Twitter, ” Proc. of WWW 2011, Hyderabad, India, March 28-April 1, 2011.

7) Y. Yamaguchi, T. Amagasa and H. Kitagawa, “Tag-based User Topic Discovery using Twitter Lists,” 2011 Int’l Conf. on Advances in Social Networks Analysis and Mining, Kaohsiung City, Taiwan, July 25-27, 2011.

8) S. Ghosh, M. Bilal Zafar, P. Bhattacharya, N. Sharma, N. Ganguly and K. P. Gummadi, “On Sampling the Wisdom of Crowds: Randow vs. Expert Sampling of the Twitter Stream”, 2013 ACM Conf. on Information and Knowledge Management, San Francisco, CA, USA, October 27th-November 1st, 2013.

9) Blei, David M.; Ng, Andrew Y.; Jordan, Michael I, "Latent Dirichlet allocation". In Lafferty, John. Journal of Machine Learning Research 3: pp. 993–1022, 2003.

10) V. Rakesh, D. Singh, B. Vinzamuri and C.K. Reddy, “Personalized Recommendation of Twitter Lists Using Cotent and Network Information”, Proc. of the 8th Int’l AAAI Conf. on Weblogs and Social Media, Ann Arbor, MI, pp 416-425, June 2-4 2014,.

11) Guanying ZHOU, Xuan Zhang and Hayato Yamana, “Identifying Topics and Influential Users based on Information Propagation in Twitter”, Proc. of SoC 2013, June 22, 2013.