KKU Engineering Journal
http://www.en.kku.ac.th/enjournal/th/

Twitter influential users ranking using Twitter user characteristics

Kanda Runapongsa Saikaew* and Wit Krutkam
Department of Computer Engineering, Faculty of Engineering, Khon Kaen University, Khon Kaen, Thailand, 40002.

Received May 2014
Accepted September 2014

Abstract

Social media sites have experienced an explosion in both the number of users and the amount of user-contributed content in recent years. There is the need for the solution for information overload in social media. In this paper, we focus on solving the problem of finding relevant Twitter users to follow and selecting only popular tweets to post, we have collected information about Twitter users, particularly the number of influential users who are followers, the number of general followers, and the number of tweets that are frequently retweeted. Then we used such statistics information to compute the user rankings. In addition, we also created a Twitter account to automatically post only tweets that have been retweeted many times. Based on the survey result and using the Spearman's rank correlation coefficient, the recommended Twitter users suggested by the system have proven to be popular and pertinent, and the rank order by the proposed system has a statistical significant degree of similarity with the user survey result.

Keywords : Twitter, Social media, Influential users, Recommendation, Ranking

1.Introduction

Nowadays, millions of people have engaged several social media apps such as Facebook, Twitter, and Google+ to keep up with their network, to participate in random chatter, to share their interests. Many active users of social network sites are nevertheless constantly troubled by information overload. There are many people to interact with and tremendous content to read. As a result, the challenge in finding the right people and content to focus on has been identified as a key challenge for social network sites.

In this paper, we focus on social media site called 'Twitter'. Twitter is a famous social media service, with millions of registers users around the globe. Twitter hosts a substantial amount of user-contributed data of real time events. Twitter core functions represent a simple social awareness stream data. Twitter users share information about upcoming events and events being attended. The users also specify their location in their profile on Twitter.

Several social networks provide APIs for programmers to collect data. Using provided Twitter API, we have developed a web application that

* Corresponding Tel.: +66-43-362-160; fax: +66-43-362-160
Email address: krunapon@kku.ac.th

recommends Twitter users, and also created a Twitter account to display only popular tweets. Although many researchers have investigated about which Twitter users should follow, most of available recommended systems consider only the number of followers. This paper has studied about other interesting parameters of Twitter user accounts, such as the number of highly retweeted tweets. In addition, we also categorize the followers as the general followers and the influential followers.

As we enter into the Twitter world, our first question is likely to be "Whom to follow?" If we follow too many users, there will be an excessive number of messages that we subscribe via our timeline. If we follow too few users, there may be some important and interesting information that we will miss. Therefore, we actually may want to follow users that are more important than others. Which Twitter user does usually provide useful and interesting content? The ability to know which users are cogent to subscribe is the motivation of this paper.

In Twitter, T_i is a follower of T_j if a Twitter user T_i follows an Twitter user account T_j and thus see the statuses of Tj on his newsfeed. In this case, T_j is a followee of T_i . If T_i is a follower of T_j and T_j is a follower of T_i, $T_i$and T_j are friends.

A tweet can contain text, emotion link or the combination of text and emotion link. Tweets have recently gained a lot of importance due to their ability to spread information rapidly. Popular search engines like Google and Bing have started including feeds from Twitter in their search results. Researchers are actively involved in analyzing these microblogging systems. Some research area attempts to understand usage and communities [1] and discover user characteristics [2].

Chen [3] studied recommendation system in microblogging which includes recommending people, recommending conversation, and recommending information. The FoF (Friend of Friend) technique introduced the old offline to be connected with friends they missed. The conversation and information recommending introduces the new friends who share common interests.

Armentano et al. [4] proposed an algorithm for recommending relevant users. The algorithm explored the topology of the network considering different factors that allowed us to identify users whom could be considered as good information sources. This algorithm first explored the target user neighborhood in searching for candidate recommendations and then sorted these candidates according to different weight features which were 1) the relation between the number of followers and the number of followees, 2) the number of occurrences of each candidate in the final list, and 3) the number of friends in common. They evaluated their algorithm by creating the survey that asked 20 students in their class. Unlike their work, our paper conducted the survey openly via the web page and asked real anonymous Twitter users to answer the survey. Then we evaluated the result by using the Spearman's rank correlation coefficient.

Wang et al. [5] proposed a scheme to detect blind sports, by contradicting explicit evidences and implicit references. The goal of this related paper is to find friends whom the user would like to interact with but may not see their updates. Based on the experimental results and surveys, their scheme could identify blind spots. This related work has done the experiments on Renren, which has functions like Facebook but is the most popular online social network service in China. On the other hand, our work focuses on Twitter, which is an online short service message. Although both this related work and our work identify users that should be followed, their

recommendations are based on personal relationships while our recommendations are based on Twitter user characteristics.

Krutkam et al. [6] analyzed the evaluation of using the number of followers and the lists of Twitter users as the ranking criteria for recommending users to follow. The experimental results showed that people tended to use the number of followers more than the number of lists. On the other hand, in this proposed work, we also consider the factor of the number of followers in the same area and the number of tweets that are highly retweeted. Guimaraes et al. [7] presented a holistic hybrid algorithm that took into account content-based, collaborative-based and user-based information. Their experiments on a real dataset from Twitter showed that their proposed hybrid method outperformed other approaches which included a baseline random algorithm, TF-IDF based algorithms, LDA based algorithms, BPR-MF algorithms, and Twittomender. Hannon et al. [8] presented Twittomender as an application for finding followees on Twitter by combining content-based and collaborative-based recommendation categories. Kim and Shim [9] proposed TWILITE, a recommendation system for Twitter using probabilistic modeling based on latent Dirichlet allocation which recommended top-K users to follow and top-K tweets to read for a user.

Kywe et al. [10] conducted a comprehensive survey to categorize the existing recommender systems into Twitter four main functions which were 1) related to tweet (what should I tweet about?, which URL should I include in my new tweet?, which hashtag should I include in my new tweet?) 2) related to retweet (which tweet should I retweet?) 3) related to mention (who should I mention in my current tweet?) 4) related to follow (who should I follow?) The authors suggested that there should be more work on hashtags, mention, or retweet recommendations. Later, Kywe et al. [11] proposed a novel hashtag recommendation method based on collaborative filtering. Their experimental results showed that their proposed method provided higher performance than the recommendation based only on tweet content. Lately, Zhou et al. [12] proposed a real-time customized recommendation scheme for microblogging systems based on the analysis of tags.

Twitter itself also recommends whom to follow. The ranking algorithm of who to follow is not revealed by Twitter. Nevertheless we observe that the recommended users have a large number of followers.Twitter also has the function called “people discover.”This function allows users to search for Twitter users with a given keyword. We have attempted to use this function with keyword “news.” The result reveals some users relevant to keyword “news” with description. However, we cannot choose the best ones to follow because Twitter does not show the ranking score. They only display the users who are relevant with the given keyword.

In this paper, there are two main contributions: 1) the recommendendation of which Twitter users to follow, and 2) the Twitter account that posts only useful tweets (assuming that useful tweets are tweets that are retweeted many times).

2. Research methodology

This section describes the details about system design, implementation, inputs, and outputs. The main concept of this section is to explain how to rank Twitter influential usersby using Twitter user characteristics.

All of the input data is collected from three methods of Twitter APIs: 1) List method, 2) Friend method, and 3) Followers method. The process unit is the portion of ranking calculation function. The output unit is a webpage that shows Twitter influential users.

The goal is to get tweets and its attributes from the feature called 'List'. First, we describe the Twitter API [http://dev.twitter.com] that is used to collect the data. In order to automatically and easily manage the collection of data obtained from Twitter, crontab has been used in our system.

We emphasize on the details of the Twitter API that provides methods to get and put information in terms of Twitter data. At the present of writing this study, there are two types of Twitter APIs: REST API, and Streaming API. The RESTREST stands for Representational State Transfer. It supports a set of HTTP methods GET, POST, PUT and DELETE to execute different operations. It executes the operation per request. The Streaming API is a global stream of real time tweet data. This paper chooses to use REST API because there are 21 methods to use for collecting data.

The REST API provided by Twitter gives access to core Twitter data. In this paper, we focus on only functions that can be collected as our system input data as following.

a) Timelines - A Twitter timeline is a list of 20 recent tweets. The provided timelines are the public timeline, friends' timelines, and user's timeline. The public timeline is a list of twenty most recent statuses of Twitter users. The user timeline for a particular user is a list of 20 most recent statuses of his friends or himself. The friends timeline is a list of 20 most recent statuses of a friend of the authenticated user.

b) Users - Users can follow other users, mark certain tweets as their favorites, tweet a status, and retweet statuses.

- User profile data: this data includes name of the user, their self-entered description, their chosen screen name, Twitter user IDs, self entered location, and URL if provided.

- Update user profile: the REST API allows access to update users profile data and options. The function requires users to be authenticated by Twitter.

c) Friends and Followers - Users follow their interests on Twitter through both one-way and mutual following relationships. Friends are the people whom the user is following. The REST API provides accesses to a list of all followers and friends for a public user.

d) Tweet – The REST API also allows posting and deleting messages to Twitter. The function requires users to be authenticated by Twitter.

e) Lists – Lists are collections of tweets, culled from a curated list of Twitter users. List timeline methods include tweets by all members of a list. Twitter provides ability to create lists of Twitter users. Users can directly follow these lists. The REST API can be used to get tweets of a list of user IDs.

The key important goals of this paper are to find the solution to overcome the information overload, in particularly to follow only the best users or the group potential users. If we follow too many users, the data stream will appear too much on our timeline.

We have implemented the system to recommend Twitter influential users by computing ranking score. We choose to focus on these three factors 1) retweets ratio, 2) followers in the same area ratio, and 3) the general followers ratio. We consider that retweets ratio is the major key factor because Twitter users usually only retweet the tweets that are worth to spread to others. Followers in the same area ratio means the ratio of whom is the most followed by other users who are in the same area. We determine whether people are in the same area by inspecting the Twitter lists that they are in. General followers ratio in this study means general Twitter users.

Typically, general users examined the number of followers to determine how famous of a given Twitter user is before deciding whether to follow a given user. Considering only the number of followers is not sufficient because we do not know whether the tweets of that user are worthwhile to follow. In addition, the ranking scores by the number of followers rarely changes while tweets are rapidly updated in real time.

If we consider only the number of followers who are in the given field as a key factor, it is actually a more appropriate indicator than the number of general followers because people in the same area are likely to know who usually give useful information. However, the ranking scores by the number of followers who are in the same area change slowly.

If we consider only the number of tweets that are frequently retweeted, the result may suggest only users that are interesting in a short period of time. Thus, this paper combines all three factors in to one equation by assigning weights to the retweets ratio, the followers in the same area ratio, and the general followers ratio.

We describe the ranking score of each Twitter user as shown in equation (1)

Ranking Score = [a*retweets ratio] + [b* followers in the same area ratio] + [c*general followers ratio] (1)

Equation description:

a is the weight for the retweets ratio, by default, it is set to 0.5

b is the weight for the followers in the same area ratio, by default, it is set to 0.25

c is the weight for the general followers ratio, by default, it is set to 0.25

The ranking score composes of three factors that are retweets ratio, followers in the same area ratio, and general followers ratio.

Retweets ratio is the number of his/her tweets are retweeted at least 20 times divided by the maximum number of tweets that are retweeted at least 20 times in a set of Twitter users within two months. The maximum period of this study for data collecting is two months. We set the retweeted threshold to be 20 times based on a preliminary experimental result that gives an appropriate number of retweeted tweets. That is, if we set the number below 20, there will be too many retweeted tweets. However, if we set the number above 20, there will be too few tweets.

Followers in the area ratio is the number of followers who are in the same area divided by the maximum number of followers who are in the same area in a set of Twitter users. Note that we update this ratio once a week. Even the followers in the same area slowly change but we should update once a week for the correct result. General followers ratio is computed by taking the number of general followers divided by the maximum number of general followers for a given user. Note that we update this ratio once a week.

We consider that the retweets ratio is the most important factor because it dynamically changes all the times. Thus the weight for retweets ratio should be more than the weight for other ratios. The followers in the same area slowly change over a long period of time. The general followers ratio slowly changes as well.

We made some experiments to find out an appropriate set of weights for the above three factors by randomly choosing a set of weights and justifying the output ranking of Twitter users with the given set of weights. Finally, we chose the set of the weights as 0.5, 0.25, and 0.25 for the retweets ratio, the followers in the same area ratio, and the general followers ratio.

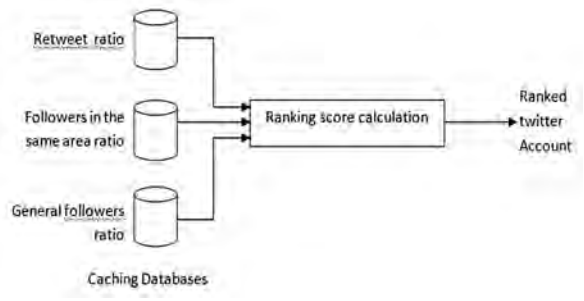

Figure 1 Ranking accounts computation

Figure 1 shows the process of users ranking computation. From this figure, the system computes the ranking score based on the number of tweets that have been retweeted at least 20 times, the number of general followers, and the number of followers in the same area.

Table 1 shows an example of some sample Twitter users statistics and how the ranking scores are computed for those users.

Table 1 Sample ranking score calculation

TU	RT	GF	FA	RScore
@a	30	276,392	282	1.00
@b	25	262,957	245	0.66
@c	18	229,782	217	0.69
@d	15	114,921	215	0.54

TA = Twitter user

RT = Retweet > 20 times

GF = General follower

FA = Followers in the same area

RScore = Ranking score

Sample Calculation

@a = [0.5*(30/30)] + [0.25*(276,392/276,392)] + [0.25*(282/282]= 1

@b= [0.5*(25/30)] + [0.25*(262,957/276,392)] + [0.25*(245/282]=0.66

@c = [0.5*(18/30)] + [0.25*(229,782/276,392)] + [0.25*(217/282]= 0.69

@d = [0.5*(15/30)] + [0.25*(114,921/276,392)] + [0.25*(215/282]=0.54

After the computation, Twitter users can be ranked as following.

Rank #1, #2, #3, and #4 are @a, @c, @b, and @d respectively

3. Research results and discussion

This section illustrates the result of the proposed system, survey results and experimental evaluation. Figure 2 illustrates the web application that recommends Twitter users in the news area. This application allows users to give weights for the retweets ratio, the followers who are in the same area ratio, and the number of general followers ratio.

Another purpose of this paper is to solve the problem of information overload. We develop the account to post the messages filtered by the proposed system to the Twitter account that we created which is called @thairanking. This account posted only tweets that had been retweeted at least 20 times. However, we could adjust the messages retweeted count to other numbers. If the retweeted count is too low, there are too many messages posted on timeline. If the retweeted count is too high, there are too few messages posted on timeline. From the observation on experiments, we have found that the retweeted count threshold as 20 times is the appropriate value of news area.

We have collected about 2,000 Twitter users and we kept only 500 news user account since the maximum number of users in the Twitter list is 500.

To evaluate the proposed system in recommending Twitter users, first, we created an online survey form in which we received the opinions from anonymous Twitter users. Secondly, we evaluated the results of proposed system to the survey result. In order to assess the proposed system, we need to know the satisfactory of the ranking

position from the anonymous Twitter users. Thus, we created an online survey form as illustrated in Figure 2.

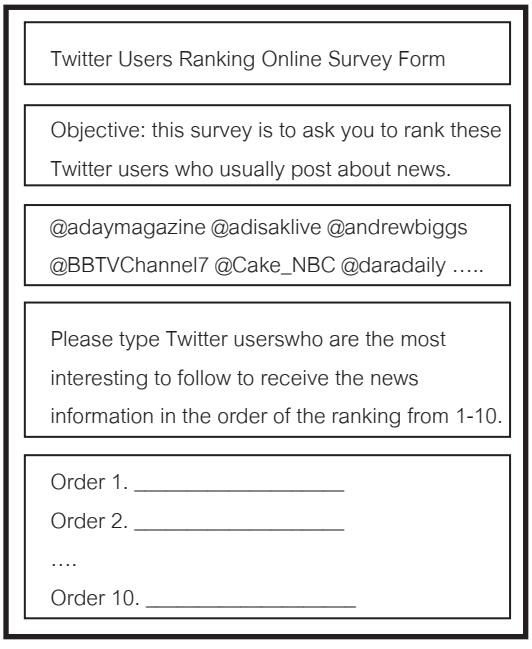

Figure 2 Questions in the survey form

The survey form essentially asked the users to rank Twitter users from the order of 1 to 10. Initially, there was a set of 50 interesting Twitter users introduced to the respondents. These users constantly posted tweet about news. To avoid bias, these 50 users are ranked by alphabets. Respondents can freely fill the provided input box, ranked from the first to the tenth by selecting from the given user or from the member in http://twitter.com/thairanking/thai-press. There were 54 respondents who filled this survey.

To evaluate the ranking of Twitter users, we use the Spearman correlation coefficient, which is defined as the Pearson correlation coefficient between the ranked variables [13]. Table 2 shows the strong positive spearman rank correlation ordered from the largest to the smallest as the rankings by the proposed system, the general followers factor, the retweet factor, and the followers in the same area factor, respectively.The strong positive spearman correlation means that there is a similar rank order between the two sets of order. The ranking order by different systems has a statistical significant degree of similarity with the survey result.

Table 2 Spearman rank correlation

Rankings by	Spearman rank correlation(r_s)
Proposed system	0.72
General followers factor	0.67
Retweet factor	0.56
Followers in the same area factor	0.52

As shown in Table 2, the ranking based on the three factors is closely to the order relative to the survey result the most. In other words, the ranking order by proposed system which is the combination by three factors (retweet, general followers, and followers in the same area) gives the better ranking order than the ranking based on each individual single factor.

4. Conclusions

In this paper, we propose a solution to recommend Twitter users by calling Twitter APIs to collect interested Twitter user characteristics, particularly the number of followers in the same area, the number of general followers, and the number of tweets that are frequently retweeted. The proposed system automatically gathered such Twitter information to recommend the potential users in real time. Based on the survey result and the Spearman rank correlation, it has been found that Twitter users suggested by the system have proven to be popular and worthwhile users in the target area. Moreover, we also developed the Twitter account that automatically

posted highly retweeted tweets, assuming that people generally wish to view social media information that many people share or give its importance.

There are some tweets that are written with different words but have the same meanings. We would like to extend our work to understand the semantics of Twitter content and group or summarize the tweets with the similar meanings. It would also be interesting to develop Twitter corpus to normalize these notations and abbreviations into the common written form.

5. References

[1] Java A, Finin T, Song X, Tsein B. Why we Twitter: understanding microblogging usage and communities. In: Joint 9th WEBKDD and 1st SNAKDD Workshop, San Jose, CA, USA: ACM. 2007, p. 56-65.

[2] Dongwoo K, Yohan J, Il-Chul M, Oh A. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In: Workshop on Microblogging at the ACM Conference on Human Factors in Computer Systems, 2010.

[3] Chen J. Personalized recommendation in social network sites. Ph.D. Dissertation, The faculty of the Graduate School of the University of Minnesota, U.S.A, 2011.

[4] Armentan MG, Godoy D, Amandi A. Topology-based recommendation of users in micro-blogging communities. Journal of Computer Science and Technology. 2012; 27, 624-34.

[5] Wang D, Liu X, Li X. Blind spots: unveiling users' true willingness in online social networks. In:Globecom 2012 – Communications Software, Services and Multimedia Symposium. 2012, p. 2066-71.

[6] Krutkam W, Saikeaw KR, Chaosakul A. Twitter accounts recommendation based on followers and lists. In: The 3rd Joint International Information & Communication Technology, Electronic and Electrical Engineering (JICTEE 2010); 2010 Dec 21-24; City of Luangprabang, Lao PDR. 2010.

[7] Guimaraes S, Ribeiro MT, Assuncao and W Meira Jr. A holistic hybrid algorithm for user recommendation on Twitter.Journal of Information and Data Management. 2013;4: 341-56.

[8] Hannon J, McCarthy K, Smyth B. Finding useful users on twitter: twittomender the followee recommender.Advances in Information Retrieval.Springer Berlin Heidelberg. 2011;784-7.

[9] Younghoon K, Shim K. TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation. Information Systems. 2014;42: 59-77.

[10] Kywe SM, Lim EP, Zhu F. A survey of recommender systems in twitter. Social Informatics. Springer Berlin Heidelberg, 2012; 7710:420-33.

[11] Kywe SM, Hoang TA, Lim EP, Zhu F. On recommending hashtags in twitter networks. Social Informatics. Springer Berlin Heidelberg, 2012;7710:337-50.

[12] Zhou X, Wu S, Chen C, Chen G, Ying S. Real-time recommendation for microblogs. Information Sciences, 2014;279:301-25.

[13] Myers JL., Well A, Lorch RF. Research design and statistical analysis. Routledge; 2010.