Mining components of micro-blogging user Influence and their correlations

Dong Liu, Quanyuan Wu, Weihong Han
School of Computer
National University of Defense Technology
Changsha, China
nudtld@yahoo.cn

***Abstract*—Measuring micro-blogging user influence is very important both in economic and social fields. In this paper, we detailedly discuss the components of user influence. Considering the time affect, *TAC*(time-effectiveness attenuation coefficient) is proposed when generating user influence which consists of post influence and follow influence. We also discuss the correlation between the two kinds of influences by use of Spearman's rank correlation coefficient. After a series of experiments, we believe that our method is more accurate and comprehensive when measuring the influences of micro-blogging users.**

Keywords-micro-blogging; user influence; Spearman's rank correlation coefficient

I. INTRODUCTION

Micro-blogging, one of the most influential media of Internet, attracts a large number of people. By the end of November 2011, the number of Chinese micro-blogging accounts has grown to 320 million and the number of tweets reached over 150 million each day [1]. Along with the overall popularization of Internet as well as the continuous development of mobile communication network, the user coverage of micro-blogging will grow much larger and its economic and social influence will continue expanding. Since more and more users acquire and share information by use of micro-blogging, mining users' features and influence has drawn many researchers' attentions

Since micro-blogging has a large number of active users, mining their influences is very helpful to measure their importance in micro-blogging. The bigger influence a user has, the more attention his behaviors draw. So the user who has big influence may affect the micro-blogging network a lot. In business recommendations, if we recommend the goods and services to the users who are more influential, others would receive the advertising information much more quirkily and efficiently. In addition, the joining of users who have big influence brings the explosive propagation of micro-blogging topics. Therefore, the analysis of user influence is much useful and essential to mine the propagation rules of topics in micro-blogging and acquire the behavior features of the key users, and it has broad application prospect in public sentiment monitoring and analyzing. Therefore, mining user influence is very useful in economic and social fields.

In this paper, we describe the components of micro-blogging user influence as well as its generation in detail. Taking Sina micro-blogging as background, a series of experiments are done to prove the effectiveness of our method.

The rest of this paper is organized as follows. Section II discusses the related work. In section III, we detailedly describe the components of user influence. Section IV gives the detailed generation of user influence. Experimental results are then discussed in section V. Finally, section VI concludes the paper.

II. RELATED WORK

Micro-blogging user influence has been studied by many researchers. Some of them analyze the user influence by use of hyperlink analysis methods, learning from PageRank [2] algorithm. Weng et al. [3] proposed the algorithm TwitterRank which ranks Twitter users. TwitterRank measures the users' influences considering the link structure of follow relationships, the similarity between users, and the number of tweets. Liu [4] proposed user-influence evaluation system called UserRank to measure the users' influence in social network based on PageRank.

The user's actions play an important role in the analysis of user influence, therefore some researchers makes comprehensive analyses considering these factors. Yuto Yamaguchi [5] focused on the post/posted, follow/followed, retweet/retweeted actions of micro-blogging users, and proposed TURank which evaluate users' authority scores in Twitter by use of a user-tweet graph based on ObjectRank [6]. Meeyoung Cha [7] did a deeply research on the users' actions and influences. He focused on three actions: follow (indegree), retweet and mention, and then he analyzed the influences represented by the three actions. At last, he measured the user influence according to the three actions, and then ranked the scores by use of Spearman's Rank Correlation Coefficient [8]. Shaozhi Ye [9] categorized the social influence to follower number influence, reply influence and retweet influence. He measured the user influence by use of Spearman's Rank Correlation Coefficient and Kendall Tau Rank Correlation Coefficient. He considered that the number of replies could reflect the user influence steadily.

These former researches consider both the users' actions and their features such as follower numbers when analyzing user influence, so the results are valuable. But there still exist some problems. For instance, these methods take no account of time affect. User influence may change as time goes by. Treating the tweets posted at different times in the same way may cause inaccurate results. Therefore, considering time

Published by Atlantis Press, Paris, France.
© the authors

affect, a comprehensive analysis of user influence is made in this paper.

III. Components of User Influence

The influence of a micro-blogging user is mainly generated from the user's features and his tweets' popularity.

The user's features mainly refer to follower number, verify information and so on. Follower number which could straightforwardly reflect the degree of user influence is the most widely used one. The more followers he has, the more widely his tweets spread and the bigger his influence is. Moreover, the other features are also very valuable in the analysis of user influence. In the paper, we mainly consider the follower number feature and the part of user influence reflects by user's follower number is called *follow influence*.

The tweets' popularity could be easily concluded from the number of users who retweet or comment the tweets. It is obviously that if a user's tweets are retweeted or commented a lot by others, he has a big influence to other users. We name the tweet's popularity *tweet influence*. A user who posts many tweets which have big tweet influences is focused by others.

Given all that, the user influence consists of tweet influence and follow influence. Besides, the time affect should be considered. A user who was influential may lose his influence as time goes by.Specifically, the time affect is represented as the changing of follower numbers and tweet's popularity.

IV. Generation of User Influence and Correlation

A. Generation of tweet influence

As described in section III, the tweet influence which represents the popularity of a tweet is generated from the number of users who retweet or comment the tweet. So, we use *retweet influence* to represent the influence reflected by retweeting number. The same is true of *comment influence*

Retweet influence of tweet t is defined as follow.

$$retweet\_influence(t)=\alpha\times retweet\_number(t) \quad (1)$$

$retweet\_number(t)$ represents the number of users who retweet tweet t. α is an adjustable parameter that show the weight of retweet influence. Generally, $\alpha\in(0, 1]$.

The same is comment influence. It is defined as follow.

$$comment\_influence(t)=\beta\times comment\_number(t) \quad (2)$$

$comment\_number(t)$ is the number of users who comment tweet t. β is an adjustable parameter that show the weight of comment influence. Generally, $\beta \in (0, 1]$. According to the experience, β is bigger than α. It means that the users who comment on the tweet t are more interested in it than others who only retweet it.

Generally, there exists a time range when mining user influence. For instance, we may want to mine a user's influence since three month ago. Therefore, we ought to consider the time which affects tweet influence a lot. It is obviously that users in micro-blogging mainly focus on the current tweets, and the tweets posted long ago may not draw their attentions. Therefore, the recent tweets have bigger influence than former ones.

TAC(time-effectiveness attenuation coefficient) is proposed to represent the attenuation degree of tweet influence. It is defined as follow.

$$TAC=\frac{1}{1+2^{(post\_time(t)-t_{now})}} \quad (3)$$

$post\_time(t)$ is the posted time of tweet t, and t_{now} is the current time. The nearer to the current time, the bigger influence a tweet has. The two time variables could be the time granularity such as 1 day which is predefined before calculating.

At last, the tweet influence of a micro-blogging user u is defined as follow. t is the tweet posted by u during the time range.

$$tweet\_influence(u)=\sum TAC\times(retweet\_influence(t)+comment\_influence(t)) \quad (4)$$

B. Generation of user influence

As described in section III, the user influence consists of tweet influence and follow influence.

Follow influence is defined as follow.

$$follow\_influence(u)=\gamma\times follower\_number(u) \quad (5)$$

$follower\_number(u)$ is the number of *followers* of micro-blogging user u, and γ is an adjustable parameter that show the weight of follow influence. Generally, $\gamma\in(0, 1]$.

Finally, The user influence is defined to be a binary vector as follow.

$$user\_influence(u)=<tweet\_influence(u), follow\_influence(u)> (6)$$

In addition, we sometimes need to mine the influence reflect by single tweet, and it means that the tweet influence sometime need to be normalized. The normalized user influence is defined as follow:

$$\begin{aligned}normalized\_user\_influence(u)&=<normalized\_tweet\_influence(u), follow\_influence(u)>\\ &=<\frac{tweet\_influence(u)}{tweet\_num},\ follow\_influence(u)> \quad (7)\end{aligned}$$

tweet_num is the total number of tweets posted by user u during the *time range*.

C. Measuring correlation between tweet influence and follow influence

In micro-blogging, the tweets of a user who has more followers always draw more attentions, so there evidently exists correlation between tweet influence and follow influence.

Rather than use the values of tweet influence and follow influence directly, we use the relative order of influence's ranks as a measure. In order to do this, we sorted users by each influence, so that the rank of 1 indicates the most influential user and increasing rank indicates a less influential user. Users with the same influence value receive the same rank.

Assume that there are n micro-blogging users, we generate tweet influence and follow influence of each u_i ($i\in$ [1,n]) at first. $tweet\_influence(u_i)$ represents the tweet influence of u_i, the same is true of $normalized\_tweet\_influence(u_i)$ and $follow\_influence(u_i)$. Then, we sort the three kinds of influence separately. The raw scores $tweet\_influence(u_i)$, $normalized\_tweet\_influence(u_i)$, *follow_* *influence*($u_i$) are finally converted to ranks *Rank_t*(u_i), *Rank_n*(u_i), *Rank_f*(u_i).

We used Spearman's rank correlation coefficient

$$\rho = 1 - \frac{6 \times \sum (x_i - y_i)^2}{n(n^2 - 1)} \qquad (8)$$

as a measure of the strength of the association between *Rank_t*(u_i) /*Rank_n*(u_i) and *Rank_f*(u_i), where x_i is *Rank_t*(u_i) or *Rank_n*(u_i) and y_i is *Rank_f*(u_i). Spearman's rank correlation coefficient is a nonparametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, the closer ρ is to +1 or -1, the stronger the likely correlation. A perfect positive correlation is +1 and a perfect negative correlation is -1.

V. Experiments

In order to ensure that the analyzed users are representative, we randomly choose 100 friends from Kai-fu Lee who is very popular in Sina micro-blogging. First, by use of OpenAPI [10], we collected the number of their followers as well as all their tweet data from October to December in 2011 include the contents, the number of comments and retweets. Their tweet influences and user influences are generated and analyzed later.

A. Analysis of user influence

First of all, we define the parameters as follows.

$\alpha \leftarrow 0.8$，$\beta \leftarrow 1$，$\gamma \leftarrow 1$ and the time granularity is defined to be 1 day.

Then the user influence is generated from the follow influence and the tweet influence by use of the formula 1 to 7, as described in section IV. Limited by the space, the result of 5 users is shown in Table I.

TABLE I. The influence of 5 users

User	*tweet_influence*	*normalized_tweet _influence*	*follow_influence*
经纬张颖	26479.90	84.06	3564527
任志强	1290319.32	278.63	12558113
邓侃	475.23	1.7	15303
李承鹏	443088.05	2080.23	6491125
作业本	725696.76	2443.42	4537586

Take "任志强" and "李承鹏" for example. The *tweet_influence* of 任志强 is much bigger than that of 李承鹏, because there are more users who retweet or comment on 任志强's tweets. However, his *normalized_tweet_influence* is smaller. The reason is that 任志强 posted 4631 tweets in the three months but 李承鹏 only post 213 tweets. So the normalized tweet influence of 李承鹏 is much higher for singer tweet.

Fig. 1 and 2 shows the distribution of the user influences and normalized user influence of the 100 users. The abscissa represents *follow_influence* and the ordinate represents *tweet_influence* (Fig. 1) or *normalized_tweet_ influence* (Fig. 2). The points positioned near the top of the diagram represent the users who have high post influence, whereas the points placed near the right boundary indicate the users who have high follow influence.

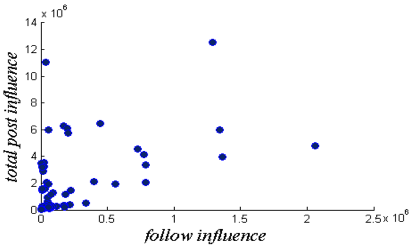

Figure 1. Distribution of user influences

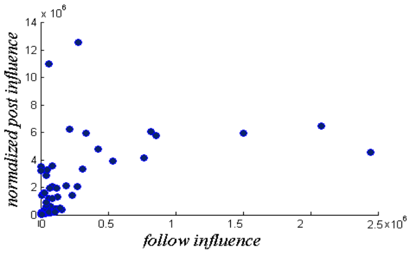

Figure 2. Distribution of normalized user influences

According to these notifications, a user has high compositive influence if the corresponding point is located near the top-right corner of the graph.

B. Correlation between tweet influence and follow influence

As described in section IV, Spearman's rank correlation coefficient is used to calculate the correlation between the ranks of tweet influence/normalized tweet influence and follow influence.

First we sort the users according the influences. Limited by the space, the result of 5 users is shown in Table II.

TABLE II. Ranks of influences of 5 users

User	*Rank_t*	*Rank_n*	*Rank_f*
经纬张颖	50	24	13
任志强	4	11	1
邓侃	83	79	85
李承鹏	10	2	3
作业本	8	1	10

The formula 8 is used to calculate the correlation, and the final result is shown as follow:

ρ*_total* = 0.7608, ρ*_normalized* = 0.8060.

ρ*_total* represents the correlation coefficient between the ranks of *tweet_influence* and *follow_influence* and ρ*_normalized* represents the correlation coefficient between the ranks of *normalized_tweet_influence* and *follow_ influence*. ρ*_total* and ρ*_normalized* both $\in$ [0,1] which means that the *tweet_influence* and *normalized_tweet_ influence* are both positive correlated with the *follow_influence*. Once the number of follower increases, the tweets draw more people's attention than before, and vice versa.

ρ*_total* < ρ*_normalized* means that the *follow_influence* influences *normalized_tweet_influence* more than *tweet_ influence*, and it also indicates that the new posted interesting tweet rather than all the tweets may attract other users' attentions.

In summary, if a user wants to improve his influence, he should post attractive tweets or increase the numbers of followers in all ways.

VI. CONCLUSIONS

Firstly, the components of user influence are described detailedly in this paper. Then, the generation of user influence which consists of tweet influence and post influence is proposed. Considering the time affect, *TAC*(time-effectiveness attenuation coefficient) is proposed as well when calculating tweet influence which consists of retweet influence and comment influence. We also discuss the correlation between the two kinds of influences by use of Spearman's rank correlation coefficient. After a series of experiments, our method is proved to be accurate and comprehensive when measuring the influences of micro-blogging users.

The methods of measuring user influence proposed in this paper are for general. Depending on different applications, such as mining the influences of users who have same topics, the methods need to be modified.

ACKNOWLEDGMENT

The authors are grateful to the anonymous referees for a careful checking and for helpful comments that improved this paper. This work was supported by National Key Technology R&D Program (No.2012BAH38B-04), "863" Program (No. 2010AA012505, 2011AA010702, 2012AA01A401、2012AA01A402) and NSFC (No. 60933005)

REFERENCES

[1] Jiang Peng, "The reduction of micro-blogging uses' activity and stability,"http://whb.news365.com.cn/jkw/201112/t20111221_3209369.htm

[2] Page Lawrence, Brin Sergey, et al., "The PageRank Citation Ranking: Bringing Order to the Web," Technical report, Stanford Digital Library Technologies Project, 1998, http://ilpubs.stanford.edu:8090/422/.

[3] Weng J., Lim E.P., Jiang J. and He Q, "Twitterrank: finding topic-sensitive influential twitterers," In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, 2010, pp. 261–270.

[4] Liu Yaoting, "Research on social network structure," Zhejiang University, 2008.6.

[5] Yuto Yamaguchi et al., "TURank: Twitter User Ranking Based on User-Tweet Graph Analysis," WISE 2010, pp. 243-246.

[6] Balmin, A., Hristidis, V. and Papakonstantinou, Y., "Objectrank: Authority-based keyword search in databases," In: Proceedings of the Thirtieth International Conference on Very Large Data Bases, vol. 30, 2004, pp. 564–575.

[7] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto and Krishna P.Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," AAAI 2010, pp. 11-13.

[8] Jerrold H. Zar, "Significance Testing of the Spearman Rank Correlation Coefficient," Journal of the American Statistical Asso, 1972, pp. 578-581.

[9] Shaozhi Ye and S. Felix Wu, "Measuring Message Propagation and Social Influence on Twitter.com," SocInfo 2010, pp. 223-228.

[10] OpenAPI of Sina micro-blog. http://open.weibo.com/