A method for measuring social influence of Micro-blog based on user operations

Mengting Hu, Hang Gao, Junming Zhou, Mingming Liu, Renhong Cheng

Abstract—Micro-blog is a social networking platform where to share short and real-time information by following mechanism. The study about social influence has attracted a number of researchers. In this paper, taking Sina Weibo as our research object, we analyze the distribution of three operations, and find the number of comment, transfer and like operation all follow power-law distribution. It is also found that the average number of three operations and the number of followers follow power-law distribution. In order to identify influential user on Sina Weibo, we define social influence of user with two aspects: user activity and user ability. User is active if user’s participation in social networking website is constantly and frequently in certain period of time. User ability is calculated by three operations, the weight of operation is assessed by AHP(Analytic Hierarchy Process). Then we propose a method named as WBRank to quantify social influence. Finally, we rank user by five kinds of assessments: three operations, the number of followers and WBRank value, and compare the correlation coefficient between every two rankings. The result demonstrates that social influence is not related with in-degree and our method of calculating WBRank value is reasonable and can find truthfully influential user.

Index Terms— Social influence, AHP, PageRank, Social Networks, Micro-blog.

I. INTRODUCTION

Micro-blogging services are prevalent all around the world in recent years. It is a convenient platform for information exchange, social communication and entertainment. In China, Weibo, short for Sina Weibo, has become the most popular Chinese micro-blog service since launched in 2009. Users can communicate, debate and express by multiple means, such as text, photo or video. Weibo has gained more than 5 hundred million registered users until 2016. According to the data published by Sina Technology in August 2016, the figure for active users has attained 2.82 hundred million monthly on Weibo.

Measuring and analyzing social influence on Weibo appeals to many researchers. Because it helps understand the value of data and monitor public opinion. According to CNNIC2(China Internet Network Information Center), users on social networking websites can be divided into four classes: (1) Entertainments stars, who possess substantial followers, expanding their influence by loyal partisanship. (2) Opinion leaders, who provide recommendations and opinions, which can stir up significant influence because of their conviction and reliability. (3) Celebrities, who are from multiple industries, such as media, business and sports. Their influence cannot be neglected as they appeal to certain group of fans. (4) The general public. They account for the largest proportion of all user and often act as the information receivers. Nevertheless, how to measure the abstract concept, user influence, is a challenging and meaningful task. Diversified subscribers and numerous number of social relationships are of great value for influence research. There are several methods to measure influence on Twitter[2, 6, 8, 11, 15, 17–19] and on Github[1], while limited studies on Sina Weibo. In this paper, we pay attention to the user operations on Weibo and measure social influence by analyzing the connection between features and business stream on Weibo. First, we find there are several issues and challenges about social influence:

• For a single micro-blog, there are three operations: comment, transfer(or retweet, in this paper, we call it transfer) and like. What is the essential operation when measuring influence? Are three of them relevant? Does the figure for operation shows some aspect of influence?

• How to find influential subscriber on Sina Weibo? How to collect data that really reflects influence.

• What measurement should be chosen to define a influential user? How to discover a solution based on these operations.

Given these problems, thus, this paper aims to discover the relationships between operations and whether operations show certain aspect of influence. Previously, the influence research concentrates on transfer and comment operation as these two operations can disseminate information to more subscribers while like operation is ignored. However, new function on Weibo makes like can broadcast information as other operations do. Therefore, like operation becomes the research priority in this paper.

The paper is organized as follows. First of all, a comprehensive preliminary is provided. It deals with the definition and measuring methods in previous research. In addition to that, PageRank algorithm and three operations on Weibo are explained. In Section 3, operations and real data on SinaWeibo are analyzed for disciplines and relations between operations. In Section 4, we propose a method to measure social influence based on PageRank. Finally, we conduct experiments and compare the ranking results among five dimensions.

Mengting Hu: postgraduate student, Nankai University, Tianjin city, China. (e-mail: humengting tongji@163.com).
Renhong Cheng: Professor, Nankai University, China.

ISBN: 978-0-9803267-9-6
II. PRELIMINARY

In this section, firstly, we retrospect previous research about social influence, which involves its definition and measuring methods. Secondly, prestige PageRank algorithm is introduced, including its principles and calculation methods, because it is prerequisite and a basic solution for this paper. Finally, three operations on Weibo are presented and each one is given a brief description.

A. Previous definition and measuring methods

For the definition of a social network influence, researchers have proposed a variety of concepts and algorithms. The Merriam-Webster dictionary defines influence as “the power or capacity of causing an effect in indirect or intangible ways.”[6]. Katz and Lazarsfeld[5] propose the opinion leaders in the two-step flow theory. The opinion leaders drive trends on behalf of the majority of ordinary people. Four core facets of influence are suggested: having a following, seen as an expert, knowledgeable expertise, and social support or social embeddedness[5]. However, in practice, influence is uncertain in today’s hybrid media environment. This means that defining influence without specific standard and data can be problematic. To address this problem, GonzalezBailon[9] proposes a certain combination of metrics, and analyses influence coverage change with time.

Who is an influential user? The wide variety of influence criteria also involves the definition of influential users. Sometimes, influential users are called opinion leaders, innovators[7] or authorities[3]. Influential users are dissemination of information and more convincing and persuasive, who lead tendency and attract attention from advertisers.

There are multiple types of methods to measure influence. The most labor-intensive method within studies of the Two-step Flow[10] is to ask people who they are influenced by and if they believe that they are influential. With large-scale social data, researchers tend to measure influence by the number of followers and/or how far a message travels[18]. However, Cha[6] find influence is not exactly related with the number of followers. Cha presents an in-depth comparison of three measuring dimensions of influence: in-degree, re-tweets, and mentions using a large amount of data collected from Twitter. They find that users with more followers may not be able to trigger more transferring and mentions. Weng[17] proposed an algorithm named as TwitterRank to find topic-sensitive influential Twitterers, which focuses on the phenomenon of homophily. Eytan Bakshy[2] finds that the largest cascades tend to be generated by users who have been influential before and who have a large number of followers.

Yu[20] studies the trending topics on Sina Weibo and discovers that trends are almost entirely due to the repeated retweets of each media content. Liao[13] analyzes user influence from three operations on Weibo: repost-only, repost and comment and comment-only, and proposes a method named WeiboRank to quantify user influence. But new feature of like operation is not taken into consideration. Liang[12] finds that in-degree and out-degree follow power-law distribution respectively, and analyzes data from multiple aspects, such as images, videos and links on Sina Weibo.

B. Basic algorithm: PageRank

PageRank[21] is, in all probability, one of the best known rank prestige methods because it underlies the Google search engine[4]. The main principle of PageRank is to track the internet Hyperlink relationship between the amount of pages, using the number of hyper-links connected to some page to calculate “PageRank value” for each page, which on behalf of the “authority” and “relevance” of the page. The algorithm first gives each page the same initial value, after a finite(in fact a relatively short) number of calculation, the result vector will gradually converge, each page will get its won PageRank value, which can be used to sort pages. Such a model is described by Equation(1) where PR(p) is the PageRank value for webpage p, M(p) is the set of web-pages linking to p and L(p) is the set of pages linked from p.

\[
PR(p_i) = \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{|L(p_j)|}
\]

In general, user has got a large number of followers, such as celebrity and star will have a higher PageRank value. On contrast, ordinary user has lower PageRank value with limited followers. If ordinary user is followed by celebrity or star, he will get higher PageRank value[8]. However, if one user only has followers but no followee, he will absorb PageRank value and calculate the result wrongly high. To solve this problem, the damping probability(usually 0.15) is add in the modified version. N is the total number of pages.

\[
PR(p_i) = \frac{d}{N} + (1-d) \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{|L(p_j)|}
\]

However, traditional PageRank method relies too much on the number of followers and distributes PageRank value equally, which can be unreasonable. We think that followee who is more influential should get more distribution, our model consider this point and will be described entirely in Section 4.

![Three operations on Sina Weibo](image)

Fig. 1. Three operations on Sina Weibo

C. Operation on Sina Weibo

On Sina Weibo, for a single micro-blog, three types of operation are provided. As shown in Fig. 1:

- Transfer: which measures through the number of reposts, same as retweet in Twitter, indicates the coverage breadth of the micro-blog.
- Like: users do the like operation by clicking the thumb symbol under the micro-blog. The number of like indicates the influence of this micro-blog in one respect.
- Comment: which measures through the number of comment,
similar to reply in Twitter, indicates the participation of the micro-blog.

In previous studies, transfer operation can disseminate micro-blog to more subscribers, thus it is regarded as the most important part of social influence. As for the comment operation, user can @ their friends to expand the influence and express their own opinions. Like operation is considered that only shows the amount but plays little part in propagating. However, when entering someone’s home page, the micro-blog that the user has clicked a like will be recommended to you, which is a new feature on Sina Weibo. Thus, the influence of like cannot only be represented by single amount. It can spread the influence the same as comment and transfer operation. In next Section, we analyze the regulation of three operations and discover the disciplines based on real data from Sina Weibo.

III. DATA ACQUISITION AND CALCULATION

A. Data Collection

At present, a great majority of data acquisition methods on Weibo make use of Sina API provided by Sina Weibo. However, Sina establishes restrictions to limit requests that user can send per hour as 150, and ordinary researcher cannot conduct a comprehensive crawling. The crawler in this experiment uses Scrapy to simulate the browser to login the Sina Weibo(http://weibo.cn), then open different pages to collect data by analyzing javascript in web page. Finally we store data to MongoDB.

In our crawler, we maintain two lists: waiting list and finished list. When crawling data, we apply breadth-first search algorithm. After collecting all data needed from a user, then remove this user from waiting list and append the user’s all followers and followees to the waiting list. Finally add the user to finished list. After crawling, four types of data are collected.

(1)User Basic Information, consist of nickname, gender, number of micro-blogs this user has posted, number of followers and number of followees. (2)Weibo, including content, user ID of poster, number of comments, number of like, number of transfer and date of the Weibo posted. (3)Follower, ID of this user’s followers. (4)Followee, ID of this user’s followees. As for the limitation of Sina, we can only get the first 20 web pages of followers and followees, and one page contains 10 members. Therefore, the maximum quantity of following relationships is 200. The basic statistics of the dataset is as follows:

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Statistics Date</td>
<td>2016.8.28-2016.9.30</td>
</tr>
<tr>
<td>Weibo Number</td>
<td>5283136</td>
</tr>
<tr>
<td>User Number</td>
<td>20987</td>
</tr>
<tr>
<td>Average Number of Weibo</td>
<td>1625.34</td>
</tr>
<tr>
<td>Number of Follower Relation</td>
<td>2059523</td>
</tr>
<tr>
<td>Number of Follower Relation</td>
<td>2899552</td>
</tr>
<tr>
<td>All Link Relationship</td>
<td>4943813</td>
</tr>
</tbody>
</table>

B. Relations between operations

Comment and transfer operation are considered as the part of influence in previous research[6, 13]. But like operation has not been studied in social influence area. To investigate the discipline, two questions arise. Question1: what is the relation and difference between like with other two operations. Question2: what are the relations between three operations with in-degree and out-degree.

![Fig. 2: Data Analysis](image)

**Fig. 2: Data Analysis**

Question 1 considers the importance of like operation. To settle this issue, at first, we randomly collect 2000 micro-blogs on Sina Weibo and calculate the number of people use these three operations. Fig. 2(a) shows the Cumulative Distribution Function (CDF) of the three operations on Sina Weibo. \( P(i) \) is the percentage of the operation \( i \) on a post. The definition of \( P(i) \) is as follows:

\[
P(i) = \frac{\text{Num}(i)}{\text{Num}(1) + \text{Num}(2) + \text{Num}(3)}
\]

\( i \) is one certain operation, \( O \) is the set of three operation, \( O = \{C,L,T\} \), \( C \) is the comment operation, \( L \) is the like operation, \( T \) is the transfer operation. According to Fig. 2(a), we find that comment and transfer operations are similar on Sina Weibo, and like operation is less than other two operations but still accounts for a large proportion. Therefore, like operation cannot be neglected when measuring influence.

Then comes to Question 2, we analyze the correlation among three operations and the number of followers for all users. The scatter plot in Fig. 2(b) demonstrates that the number of three operations are positive correlation with in-degree in logarithmic space, which predicates that some user has more followers, his or her micro-blogs possess more average number of three operations. In addition to that, the figure for three operations and in-degree all follows power-law distribution. As for the disciplines of three operations, we assume the number of operations follows power-law distribution and then fit curve.
At first, the cumulative probability should be calculated.

\[ cp = \log \frac{op_{num} + 1}{\max(op)} \] (4)

Where \( cp \) is the cumulative probability of \( op \), \( op_{num} \) is the number of samples of \( op \), \( \max(op) \) is the max number of \( op \). Fig. 3 displays the fitting curve of three operations.

\[ f_c(i) = x^{0.53} \cdot e^{-3.57e-05} \] is the cumulative probability fitting curve function of comment operation. \( f_r(i) = x^{-0.46} \cdot e^{-7.42e-06} \) is the cumulative probability fitting curve function of transfer operation. \( f_l(i) = x^{-0.43} \cdot e^{-7.97e-06} \) is the cumulative probability fitting curve function of like operation. The function is.

The derivation of every function represents the probability distribution density, which is the ability of single operation of user \( i \). According to Pickering, if cumulative frequency graph plotted on log axes, a straight line would indicate a good fit to a power-law model[14]. Therefore the result demonstrates three operations follow power-law distribution. In addition, there is no relation between operations and out-degree, which signifies that out-degree cannot contribute to social influence.

IV. MODEL

The Weibo network contains three elements \( G = (V,E,W) \), where \( V \) is the set of users, any \( m \in V \) is a particular Weibo user. \( E \) is the directed edge set. \( e = (m,n) \in E \) represents user \( m \) follows user \( n \). Meanwhile, if \( (n,m) \in E \), user \( n \) follows \( m \). \( W \) is the set of weight, any \( w_{m,n} \in W \) represents the weight in edge \( (m,n) \). Based on the analysis in last section, the weight of edge is defined as:

\[ w_{m,n} = A(n) \cdot I_{op}(n) \] (5)

\( w_{m,n} \) is the weight of edge of \( (m,n) \). \( A(n) \) is the activity of user \( n \). \( I_{op}(n) \) is the ability based on operation of user \( n \). Then specific descriptions are demonstrated in next subsection.

A. Evaluation Factor

The weight of edge \( (m,n) \) represents the proportion of social influence that follower \( m \) allocates to followee \( n \), which indicates that among all followees of user \( m \), the more powerful of user \( n \), the more social influence value that flows from \( m \) to \( n \). In our model, user’s power is evaluated by user activity and user ability. The original two subsections below describe these two factors and the third subsection demonstrates how to calculate user ability.

1) User activity: In the message mechanism of Sina Weibo, micro-blog is automatically recommended to the followers’ message page. Thus, the more user posts Weibo, the more user can influence others. Where \( A(n) \) is the activity of user \( n \). Users are active when their participation in social networking website is constant and frequent in a period of time, regardless of the attention they receive for their participation[15]. User activity refers to the average number of micro-blogs this user publishes in a period of time, the expression is:

\[ A(n) = \frac{Weibo_{num}}{T} \] (6)

where \( Weibo_{num} \) is the number of micro-blogs published by user \( n \) in the period of \( T \). Therefore, Equation (6) aims to compute the average number of micro-blogs that user \( n \) releases during period \( T \).

2) User ability: User ability represents the ability when he is recognized by other users in social networking website. More quantities of comment, transfer and like indicate that the user has stronger ability and his micro-blog is more powerful. According to the analyses in Section 3, we find the figure for transfer, like and comment follow power-law distribution respectively. When quantifying social influence, three operations represent user ability, which can be calculated by Equation (7).

\[ I_{op}(n) = \lambda \cdot C_{num}(n) + \beta \cdot T_{num}(n) + \gamma \cdot L_{num}(n) \] (7)

\( C_{num}(n) \) is the average number of comment of micro-blogs published by user \( n \). \( \lambda \) is the weight of comment operation. \( T_{num}(n) \) is the average number of transfer of micro-blogs published by user \( n \). \( \beta \) is the weight of transfer operation. \( L_{num}(n) \) is the average number of like of micro-blogs published by user \( n \). \( \gamma \) is the weight of like operation. The method to calculate the weight of three operations is demonstrated below.

3) Calculate coefficient weight based on AHP: AHP(Analytic Hierarchy Process) is a decision-aiding method developed by Saaty[16], which quantifies relative priorities for a given set of alternatives. We need to establish a hierarchical model and construct the judgment matrix. Then, calculate the largest eigenvalue and the corresponding eigenvector. The result is as Table II shows:
After calculation, the consistency ratio is 0.002, which is less than 0.1. Therefore, the result is reasonable and compliance with the consistency check.

B. Model based on PageRank

Based on the principle of PageRank, in our model, user influence of n is evaluated by a specific value, WBRank(n) value, which is the summation of WBRank value that all followers of user n that allocates to user n.

\[ \text{WBRank}(n) = \sum_{m \in F_r(n)} \text{W}(m, n) \cdot \text{WBRank}(m) \quad (8) \]

\( \text{W}(m, n) \) is the proportion of WBRank value that user m allocates to the followee n. \( F_r(n) \) is the followers set of user n.

When WBRank value is flowing in the graph, there exists probability that it encounters dead end, which indicates that the WBRank value only transfers in some circle and cannot contribute to other nodes. In case of the dead end, \( d_e \) is appended to Equation (8) and it is the possibility that the flowing direction jumps out of the circulation. Therefore, Equation (8) is updated and WBRank value is calculated by Equation (9), Set \( d_w = 0.85 \).

\[ \text{WBRank}(n) = (1 - d_w) + d_w \sum_{m \in F_r(n)} \text{W}(m, n) \cdot \text{WBRank}(m) \quad (9) \]

\( \text{W}(m, n) \) is the proportion of user m allocating WBRank value to followee n, which can be calculated by Equation (10). In this Equation, \( w_{m,n} \) is the weight of edge \((m, n)\) and \( F_r(m) \) is the followee set of user m. This equation is computing the proportion that user n accounts for among all followee of user m, which represents the percentage of WBRank value that user m distributes to n.

\[ \text{W}(m, n) = \sum_{i \in F_r(m)} \frac{w_{m,i}}{w_{m,n}} \quad (10) \]

When calculating final result, the process can be transformed to matrix computation in Equation (11).

\[ \text{WBRank} = (1 - d_w) + M \cdot \text{WBRank} \quad (11) \]

\( \text{WBRank} \) and \( (1 - d_w)M \) is a N-dimensional vector. \( M \) is a \( n \times n \) two-dimensional matrix, and \( n \) is the number of users.

The definition of \( M \) is: \( M_{mn} \) is \( WI(m, n) \) if user m follows user n, else \( M_{mn} = 0 \). After finite times of iterative computations, WBRank vector will gradually converge and each user can get his or her own WBRank value, which is the outcome of quantifying social influence and this value can be applied to rank.

V. EXPERIMENT

A. Data Preprocessing

Before the experimental calculation, dataset is massive and some information is unavailable. Thus the information should be filtered left available and valuable data. We conduct the following data preprocessing: (1) Select ponderable user. We filter left 15481 users who publish at least one micro-blog from January 1st, 2016 to June 1st, 2016. Because if the user has not release one micro-blog in this period of time, his or her user activity is zero in our model, which is meaningless in our method. (2) Sort out valuable micro-blogs. Due to time dimension is considered in our model, micro-blogs are valuable if they are published in first half year of 2016, which are filtered left 1492251 records. (3) Filter user relationships. Because the dataset contains more user relationships that information of one or two users cannot be crawled. Therefore, we filter left 2534354 user relationships that both user are in the user set.

B. Experimental evaluation

At first, we calculate WBRank value for each available user and rank all users by the amount of their value. To contrast with other methods, we rank the users by four other dimensions. The ranking results are as Table III shows:

Table III: Top 10 user ID by different rank methods

<table>
<thead>
<tr>
<th>Comment</th>
<th>Like</th>
<th>Transfer</th>
<th>In-degree</th>
<th>WBRank value</th>
<th>Sina Influence Billboard</th>
</tr>
</thead>
<tbody>
<tr>
<td>1259110474</td>
<td>1259110474</td>
<td>1259110474</td>
<td>2671109275</td>
<td>2671109275</td>
<td>88</td>
</tr>
<tr>
<td>5563960339</td>
<td>5780931227</td>
<td>5780931227</td>
<td>1259110474</td>
<td>1259110474</td>
<td>130</td>
</tr>
<tr>
<td>2289674265</td>
<td>1648764935</td>
<td>2289674265</td>
<td>1642512402</td>
<td>2590247225</td>
<td>378</td>
</tr>
<tr>
<td>1774978073</td>
<td>5902696506</td>
<td>1774978073</td>
<td>1738932247</td>
<td>2534354</td>
<td>704</td>
</tr>
<tr>
<td>2845720390</td>
<td>2874933414</td>
<td>2874933414</td>
<td>1787920531</td>
<td>1787920531</td>
<td>468</td>
</tr>
<tr>
<td>2634154091</td>
<td>2289674265</td>
<td>2562974225</td>
<td>2886747265</td>
<td>1259110474</td>
<td>720</td>
</tr>
<tr>
<td>5780931227</td>
<td>2634154091</td>
<td>1791811587</td>
<td>2289674265</td>
<td>2845720390</td>
<td>7322</td>
</tr>
<tr>
<td>1648764935</td>
<td>5619426492</td>
<td>1648764935</td>
<td>2634242155</td>
<td>1738932247</td>
<td>2944</td>
</tr>
<tr>
<td>1290404267</td>
<td>2671109275</td>
<td>2634154091</td>
<td>216813703</td>
<td>2289674265</td>
<td>7317</td>
</tr>
<tr>
<td>5902696506</td>
<td>1774978073</td>
<td>2874933414</td>
<td>1787920531</td>
<td>5902696506</td>
<td>9001</td>
</tr>
</tbody>
</table>
(1) More number of three operations does not indicate higher WBRank value and more influence power. ID = 1259110474 is in the first place under three dimensions, consist of comment, like, transfer. The user is a famous actress in China and the number of her followers is more than forty million. In WBRank value column, this user is in sixth place. Obviously, five users are more influential than her ranking by our method. Compared with ID = 1259110474, micro-blogs of these five subscribers have fewer number of comment, like and transfer. However, these five users are more active and constantly participate in social networking website. That’s the reason why their WBRank value are higher. On contrast, ID = 2671109275 is in the first place under WBRank value dimension and it is under 88th place under Sina Influence Billboard, which is also the most influential user in our dataset. This user has more than ten million followers, which contributes to his influence power.

(2) WBRank model can discover really influential users. In Table III, comparing last two columns, eight in ten users have the same sequence. For example, ID = 2671109275 has the highest WBRank value, who ranks eighty-eighth on Sina Influence Billboard and he is the most influential among ten uses in WBRank value column as well. The rate of WBRank model is probably 80% among top ten users.

Therefore, measuring social influence by any single dimension is not reasonable. The result of WBRank is integrated by multiple dimensions, which has no problems mentioned before. And the experiment shows that the most influential and valuable user is found by our model.

VI. CONCLUSION

This paper analyze the social influence of Sina Weibo by employing three operations: comment, transfer and like. We crawl real dataset from Sina Weibo by the web crawling framework Scrapy. Firstly, we analyze the importance of three operations by CDF(Cumulative Distribution Function) and found the importance of transfer and comment are common, while like operation is less important. Secondly, we analyze the distribution of three operations and find that they all follow power-law distribution. Then, for every user, we calculate the average number of three operations, and study the relation between operations and in-degree and out-degree. We find that three operations are positive correlation with in-degree in the logarithmic space. Therefore, the number of three operations and in-degree follow power-law distribution. We also find there is no obvious relation between three operations and out-degree.

Focusing on the relationships between three operations, we propose a model named as WBRank that is based on PageRank to quantify the social influence. The weight of three operations is assessed by AHP. We rank users by five methods and compare the ranking list of each dimension. The results prove that our model is reasonable when discovering the truthfully influential user on Sina Weibo. And social influence is not much related with in-degree, which is same as the study of Cha[6]. The biggest contribution of this paper is that we find three operations on Sina Weibo all follow power-law distribution and we propose a model that can discover really influential users on Weibo.

VII. FUTURE WORK

In our future work, we would like to conduct an insightful research on new operations on Sina Weibo, such as closing comment and setting permissions. Afterwards, we will concentrate on the aspect of information stream to study social influence, the stream of information may shows some respect of social influence.

REFERENCES