User Influence Analysis Model for Weibo Topics

Guixian Xu *, Yuan Tian, Yueting Meng

* Corresponding author: guixian_xu@muc.edu.cn

School of Information Engineering, Minzu University of China, Beijing 100081, China

Abstract: Due to the openness of social media, public opinion events are often triggered. Identifying important users in hot topics is helpful to correctly guide public opinion and create a green online environment. Directed at the fact that the existing methods fail to consider the influence of users' followers and the influence of comment sentiment tendencies, a user influence analysis model for Weibo based on user information and content information - UCRank (user-content influence rank) - was proposed. The model takes into account four factors, users self-influence, followers-influence, content information and comment sentiment polarity, to jointly quantify the user influence. Experimental results show that the proposed model has the best performance in terms of precision, recall and F1 value when compared with other traditional algorithms.

Keywords: Social network, Weibo topics, user influence, emotional tendencies, PageRank

1. INTRODUCTION

With the advent of the Internet era, more and more people have started to express views and emotions on hot events through social networking platforms such as Twitter and Sina Weibo. As the most popular social media, Weibo has become an important window for internet users to obtain information, spread their opinions and share their views. Famous users such as government departments, celebrities or "Big V" on the Weibo usually have a large number of followers, and their comments on social media often spread quickly. These users often play an important role in the spread of topics[1]. Modelling and estimating the influence characteristics of users is important for opinion information management[2], research to discover the pattern of microblog topic dissemination[3], and microblog friend recommendation[4].

In recent years how to accurately quantify user influence through big data mining has become one of the hot issues in academia nowadays, and many methods of influence calculation have been proposed. Shi Lei et al[5] constructed a user activity model based on three behaviors: retweeting, commenting and @, taking into account the distribution and messaging characteristics of different users, but the method did not consider the relationship between users in the social network. Kwak et al[6] analyzed the influence of users based on the number of followers and the PageRank algorithm, based on the network relationship between users, but ignored the nodes' own attributes and the content of blogs. Wu Hui et al[7] combined the content of users' tweets with the network topology to calculate user influence, but this method focused on the quality of users' blog posts but not on the sentimentality of the content.

Most current analyses of user influence focus on user social networks and user behavior, and most research results focus on measuring the degree of influence of a user on other users, with

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334300

little research on how users are influenced; moreover, the quality of content information and comment sentiment polarity are neglected. Therefore, this paper proposes a user influence analysis model, UCRank (user-content influence rank), which integrates user information and content information, taking into account user social structure, user behavioral characteristics, blog post quality and comment sentiment tendency. The proposed method has the best performance in user influence analysis when compared with the leading methods of IndegreeRank, TwitterRank and PageRank.

2. RELATED WORK

User influence is characterized by the ability to elicit potential behavior from others and to effectively communicate information, and Weibo user influence greatly reflects the actual social influence of users. There are currently three main methods used to evaluate user influence: methods based on network topology, methods based on web links and methods based on user behavioral characteristics.

Early approaches were based on network topology to evaluate user influence. The social network analysis approach considers networks as consisting of many nodes with dependencies and collaborative relationships [8], the result of people's interactions in social networks, and the most direct source of data for analyzing influence. Hao et al[9] analyzed and proposed a CSSM algorithm that considers outward centrality and neighborhood to calculate node impact. However, the algorithm ignored the contribution of tweets to user impact. Pei et al[10] divided nodes into subtypes and proposed a scalable SIIE-h method to estimate individual impact, and demonstrated the accuracy and robustness of the method to network dynamics. In addition, influence maximization[11-13] is one of the methods based on the network structure. However, in actual social networks, nodes with a higher degree of centrality or intermediacy do not always have greater influence, while assessments that consider only the network structure ignore the role of users' information in the process of social network occurrence and lack semantic interpretability and explanation of user behavior.

To effectively measure user influence, Google founders SergeyBrin and LawrencePage proposed the PageRank algorithm in 1998, drawing from the classical web ranking model algorithm. The algorithm considers not only the degree of entry, but also the importance of node neighbors, which can be recursively transferred to other nodes. In recent years, researchers have proposed many more improved page rank algorithms. Boyd et al[14] studied the actual situation of user retweets based on PageRank and proposed a Twitter User Rank algorithm that combines multiple ways to study user behavior such as user identity, tags and communication loyalty. Drawing on this idea, Lijun et al[15] re-reviewed the current Chinese Sina Weibo influence research algorithm and proposed the Weibo User Rank algorithm for the specific case of Weibo. The method based on web link analysis was first used to measure the influence of web pages and achieved better results in web page ranking, and later improved to measure the influence of microblog users. However, it is directly used to analyze user influence with more defects.

User behavior is one of the most useful features of online social media, and is the basis for studying user behavior, information dissemination models, and is important for user influence analysis. Therefore, people start to study users' social influence from their historical interaction behavior records. Cano et al[16] proposed a retweet subgraph based on Twitter graph, and used the topic-entity relevance of retweeting relationships to analyze the influence of users' topics and entities to discover influential users. Rezaie et al[17] combined user behavior and configuration, and proposed an influence index cumulative index and mean index to achieve the identification and ranking of high-influential users after a specific event, however, the method lacks the judgment of influence polarity. In response to the situation of harmful information dissemination in social networks, Li et al[18] proposed an influence calculation method PUI based on user text and behavioral features, using gradient-enhanced decision trees to classify users and analyze their influence ability in the process of information dissemination. The evaluation method based on user characteristics only analyzes the information of users' own attributes and does not consider the social structure among users in social networks and does not fully absorb the detailed information of the interaction process. Therefore, there is still a lack of comprehensive studies that measure the key characteristics of user influence.

Previous work has focused on user factors such as social structure and user behavior to measure the impact of users on social platforms. However, in addition to human factors, both the quality and content of blog posts may play an important role in enhancing user influence. To tap this influence pattern, this paper takes Weibo as the research object, conducts an in-depth analysis of user influence in hot events, and proposes a Weibo user influence analysis algorithm that integrates user information and content information. Firstly, we analyze the user's own information and build a network model integrating multiple perspectives from four aspects: user's prestige value, user's activity, content‘s spread ability and innovation, in order to more accurately describe the user's own influence; secondly, we analyze the quality of user's followers and draw on the idea of PageRank algorithm to get the indirect influence based on user's relationship; then, based on microblog content information and microblog comment information, we integrate users' emotional tendency, analyze and judge users' opinions and attitudes on topics from the semantics of texts, and obtain content influence based on microblog content; finally, we combine the two main factors of user influence and content influence, and integrate user interaction strength factor to jointly quantify the final user influence.

3. USER INFLUENCE ANALYSIS MODEL

In this paper, we analyze user influence in hot events in four aspects: users self-influence, followers-influence, content information and comment sentiment polarity, and propose an analysis algorithm that integrates two main factors, user information and content information, to jointly quantify microblog user influence. The schematic diagram of the overall user influence analysis model is shown in Figure 1.

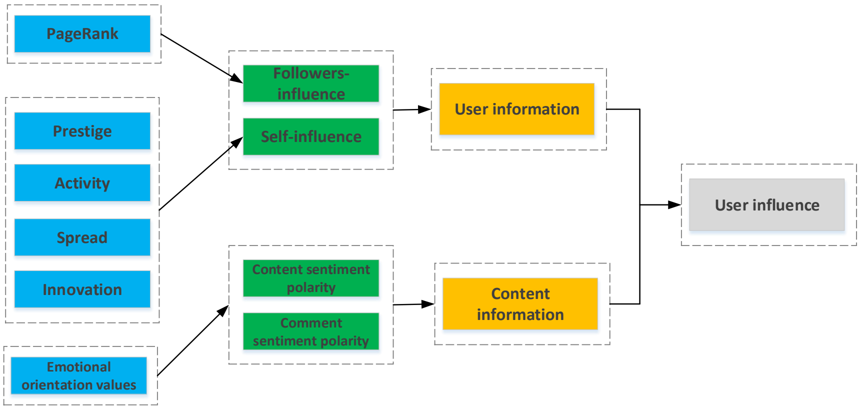

Figure 1. Schematic diagram of the user influence analysis model proposed in this paper

3.1 User influence based on user information

(1) User's self-influence calculation

In the Weibo platform, the factors that indicate whether the user has enough influence are: whether the user has fame in the social platform(Prestige); Whether the user is active enough(Active); Whether the communication power of blog posts published by users is strong enough(Spread); Whether the blog posts published by users are original(Innovation). so we combine these four factors to calculate the users' own influence:

$$INF_{self}(U_i) = a * \frac{P(U_i)}{maxP(U_i)} + b * \frac{S(U_i)}{maxS(U_i)} + c * \frac{I(U_i)}{maxI(U_i)} + d * \frac{A(U_i)}{maxA(U_i)} \quad (1)$$

Where P, A, S and I are prestige, activity, spread and innovation of the user's self-influence, which can be quantitatively calculated. However, the importance of a, b, c, d can only be qualitatively analyzed. Based on previous literature studies and parameter analysis, we consider the parameter importance: prestige > spread > innovation > activity, and refer to the setting of parameters in Ouyang's paper[19], the weight factor is set as a=0.4, b=0.3, c=0.2, d=0.1 in this paper.

(2) User's followers-influence calculation

The fans Index can be used to assess the quality of the fans a user has, and we have observed that if a user has a large number of high-impact fans, then that user is also very likely to be a high-impact user. Therefore, the quality of users' fans is also an important indicator of user impact. In order to make full use of the information of users' fans, we use the number of users' fans to measure the quality of their fans, at the same time referring to PageRank algorithm, users' fans are interested in chaining out the web page, users' fans are chaining into the web page. Get the influence of your end-user fans as shown in formula (2):

$$INF_{fans}(U_i) = (1 - q) + q * \sum \frac{INF_{fans}(U_j)}{F(U_j)} \quad (2)$$

Where U_j is the fan of the user U_i, $F(U_j)$ is the number of fans of the U_j, q is the damping factor, 0.85 in this paper.

(3) Influence analysis based on user information

This paper puts forward a research which combines the influence of users and fans. It uses both user relationship to get direct influence and user fan index to get indirect influence. The specific definition is shown in formula (3):

$$INF_{user}(U_i) = w * INF_{self}(U_i) + (1-w) * INF_{fans}(U_i) = w * INF_{self}(U_i) + (1-w) * \frac{1}{100}\sum_{j=1}^{100} INF_{fans}(U_j) \tag{3}$$

In Weibo, although personal influence is positively related to the corresponding indicator, when the indicator approaches infinity, influence tends to a fixed value, that is, influence eventually converges. Therefore, with regard to the impact of selection on fans, we arbitrarily select 100 of them and take their average value as the indirect influence of users. According to previous literature research and parameter analysis, the weight factor of user influence and fan influence is set to: w=0.75.

3.2 User influence based on content information

The core of the impact of microblog users is the effective transmission of content information. The quality of microblog content published by users will directly affect the impact of this user[20]. The basic attribute indicators of microblog include the content of microblog, the publishing time, the average length of replies, and the emotional guidance value of comments. Among them, microblog content refers to the information contained in the body of the microblog, including pictures, videos, @symbols, topics and content length. Define $C1(U_i)$, $C2(U_i)$, $C3(U_i)$, $C4(U_i)$ as whether the content of the microblog contains topics, pictures, videos, @. If it contains, then set the value to 1, otherwise 0.

And define Len1 (U_i) for the length of the microblog content and Len2 (U_i) for the average length of all comments received by the microblog text.

In the dissemination of hot events, users are the core of social media, and their negative emotions are the important characteristics of public opinion dissemination on social networks. Experience has shown that microblog comments with negative emotional tendencies can be more emotional to other users, causing events to spread widely, leading to the outbreak of hot events, and users will have greater emotional influence. Therefore, this paper introduces emotional orientation values to measure the negative emotional tendency of users' microblog comments in hot events, and the calculation method is shown in Equation (4):

$$Senti(U_i) = \frac{neg-pos}{neg+pos} \tag{4}$$

Where $Senti(U_i)$ is the emotional orientation values of Weibo text, which ranges from (-1,1). The closer the value is to 1, the more users who have negative views on their content, the more likely they are to cause hot spots. The neg is the number of comments with negative emotions in the Weibo text, pos is the number of comments with positive emotions in the Weibo text.

This paper presents a sentiment feature analysis model combining local and global features to calculate the sentiment polarity of comments. As shown in Figure 2, the text is first converted to a vector representation by word2vec model. For the local word vector features, the local word vector features are input into the BiLSTM model for training. For the global document theme feature, this paper introduces the theme model into feature extraction, fuses the neural theme model to expand the text feature, and uses the text theme feature as the global feature to represent the text information. Then the text features based on local weighted word vector and neural theme model are stitched to get text feature vectors containing subject information. Finally, the stitched text vectors are output using softmax layer to complete user sentiment feature analysis.

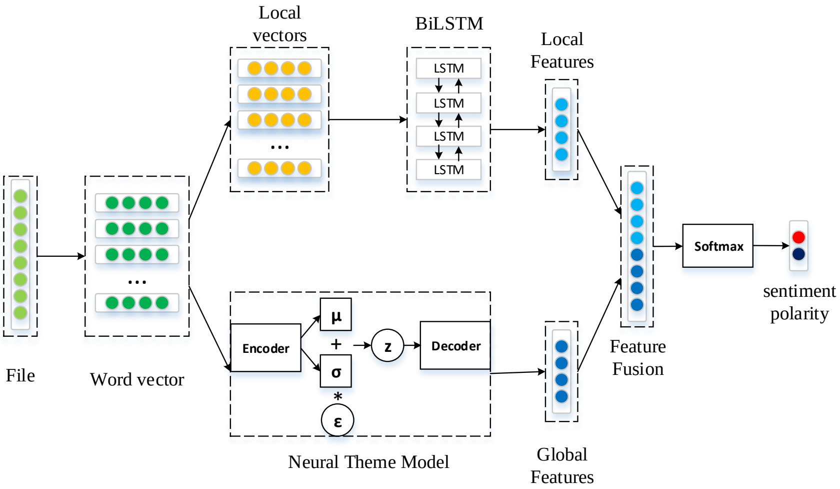

Figure 2. Schematic diagram of the sentiment analysis model proposed in this paper

To sum up, the user influence based on content $INF_{content}$ is shown in Equation (5):

$$INF_{content}(U_i) = \frac{1}{7}\left(C1(U_i) + C2(U_i) + C3(U_i) + C4(U_i) + \frac{Len1(U_i)}{maxLen1(U_i)} + \frac{Len2(U_i)}{maxLen2(U_i)} + \frac{Senti(U_i)}{maxSenti(U_i)}\right) \quad (5)$$

3.3 User influence based on user information and content information

Based on the above two dimensions of user influence based on user information and content information based on microblog content, the intensity factor is incorporated to quantify the user impact of microblog. To sum up, then can get the final INF (U_i) representation of Weibo user influence as shown in Equation (6):

$$INF(U_i) = w_1 * INF_{user}(U_i) + w_2 * INF_{conten}(U_i) \quad (6)$$

Where w_1, w_2 is the adjustment factor. According to previous literature research and parameter analysis, the weight factor of user's self-influence and followers-influence is set to: w_1=0.6, w_2 = 0.4, balancing the effects of independent variables.

4. EXPERIMENT

4.1 Dataset

This paper collects hot topics on Sina's official microblog titles "Chengdu female driver beaten" and "Jiuzhaigou earthquake". For the raw data collected, the data is preprocessed, using the number of days from registration to January 1, 2022 as the registration time. In order to reduce the interference of zombie powder, users with fewer than 20 fans and less than 20 blogs are excluded. The final number of users captured was 11201, and the total number of speeches reached 138091.

4.2 Comparison Model

In order to verify the effectiveness of the method, the classical user influence analysis algorithm or the current more popular algorithm is selected to compare with the algorithm in this paper. They are as follows.

(1) IndegreeRank[20]: the number of followers of users in social networks is used to evaluate user influence.

(2) TweetRank[21]: the number of blog posts posted by users in social networks is used to evaluate the user influence.

(3) PageRank: the method quantifies the influence of nodes based on the social network topology only, and calculates the influence by the number and importance of the user's followers. In this case, the damping factor d = 0.85 is set, and the convergence error is set to 0.001.

4.3 Experimental evaluation method

In this paper, the UCRank and the above three algorithms are cross-validated to determine the ranking of real user influence. Let U_A, U_B, U_C and $U_D$be the set of Top-k users calculated by IndegreeRank, TweetRank, PageRank and UCRank, respectively, and the set of reference users is U_R. The formula for calculating the set U_R is as follows (7):

$$U_R = (U_A \cap U_B) \cup (U_A \cap U_C) \cup (U_A \cap U_D) \cup (U_B \cap U_C) \cup (U_B \cap U_D) \cup (U_C \cap U_D) \qquad (7)$$

Precision is used to measure the checking accuracy in classification, which refers to the authenticity of the users with high influence identified from the dataset. The precision rate P_A calculation formula is shown in equation (8).

$$\mathrm{P_A} = \frac{|(\mathrm{U_A} \cap \mathrm{U_R})|}{\mathrm{U_A}} * 100\% \qquad (8)$$

Recall is used to measure the check-all rate in classification and is used to verify the ability of the algorithm to find users with high influence. The recall rate $\mathrm{R_A}$ is calculated as shown in equation (9).

$$R_A = \frac{|(U_A \cap U_R)|}{U_R} * 100\% \qquad (9)$$

The F1 score is used to combine the precision and recall of the model, and F_A is calculated as shown in equation (10).

$$F_1 = \frac{2 * P_A * R_A}{P_A + R_A} * 100\% \qquad (10)$$

4.4 Analysis of experimental results

(1) Comparison experiments of different models

In order to obtain more accurate evaluation results, the top 500 users of influence ranking were selected as the key research objects. The set of users whose user ranking is in Top-500 is determined by at least three models as the experimental reference set, and their experimental effects are calculated. the experimental results of Precision, Recall, and F1 are shown in Figure 3, 4, and 5, respectively.

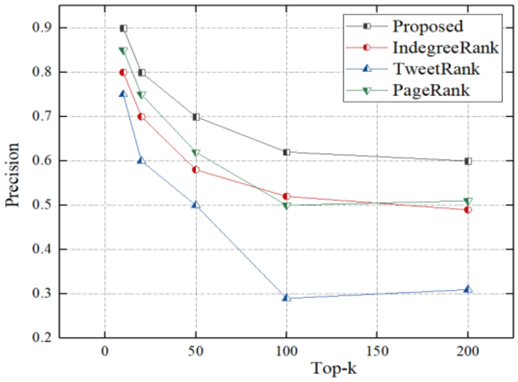

Figure 3. The precision of different algorithms among Top-k users

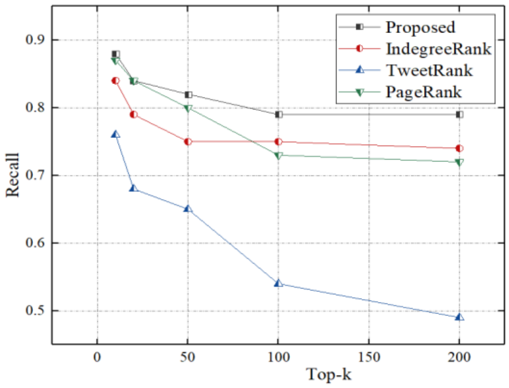

Figure 4. The recall rate of different algorithms among Top-k users

As the experimental results in Figure 3 show, the precision of the UCRank achieves optimal results when selecting Top-k users of different sizes for algorithm precision analysis, and the precision gap between the UCRank and the comparison models shows a tendency to gradually increase as the number of Top-k users gradually increases, which indicates that the UCRank is significantly better than the comparison models and can more effectively evaluate the size of the user's real influence.

The experiments also compare the recall of several models at different user sizes, and it is clear from the experimental results that the recall of all the models in this paper is higher. When k is small, the recall of all models is not much different, but as k increases, the recall of other models is significantly lower than that of the UCRank. And when k is small, the larger the average recall is, the greater the practical significance. This result indicates that the UCRank obtains more accurate influence ranking because it takes into account the user's self-influence, indirect influence of followers and influence of microblog content.

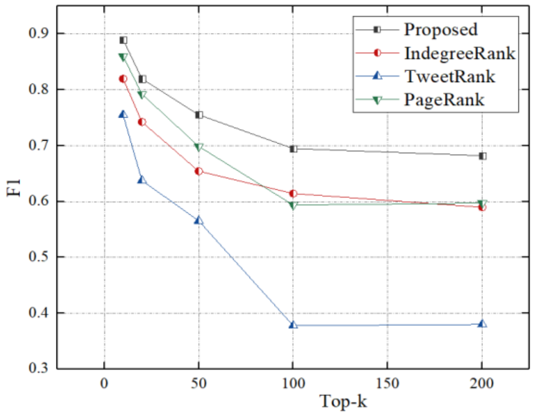

Figure 5. F1 score of different algorithms among Top-k users

From the F1 of different algorithms, we can see that the F1 values of all the algorithms in this paper are higher than those of the other algorithms. This indicates that the algorithm proposed in this paper can, to a certain extent, prevent users from using false behaviors such as increasing the number of "zombie" followers and publishing spam blog posts to maliciously increase their influence.

(2) User influence profiling experiment

Each user in the dataset is calculated by four algorithms for user influence, and the users are ranked in order of influence value from largest to smallest. In order to more easily analyze the ranking of the UCRank proposed in this paper, the number of followers and tweets of Top10 users are shown in Figures 6 and 7.

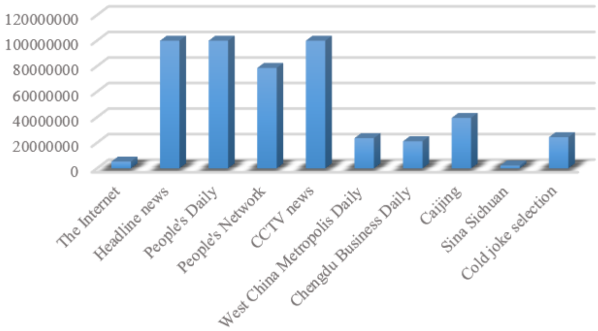

Figure 6. the number of followers of the top 10 users

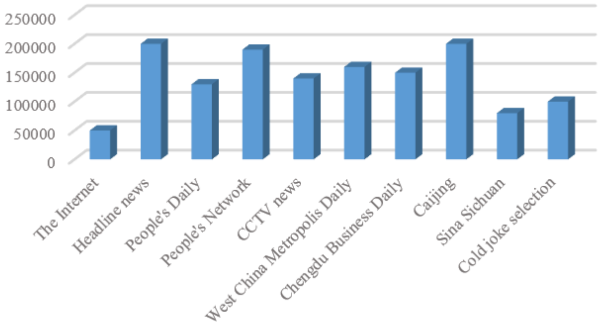

Figure 7. the number of tweets of the top 10 users

From the bar chart of the number of followers, we can clearly see that UCRank's Top10 user ranking does not form a positive correlation with the number of followers, which can be concluded that users with a large number of followers do not necessarily have a high influence. For example, the user " The Internet " did not enter the top 10 in the three comparison algorithms, but it ranked first in UCRank. After an in-depth study, we found that although the number of followers of this user is small, the number of retweets, comments and likes of this user's microblog in this hot event is very large, which means that this user interacts a lot with other users and has a strong influence. This indicates that the user has a strong ability to disseminate information by interacting with other users. In summary, although the number of followers is an important reference index of users' influence, it should also be appropriately combined with other factors. As shown by the comparison of the bar chart of the number of microblogs, users with more blog posts do not necessarily have more influence. For example, the user ranked No. 1 has a high number of likes, comments and retweets on his blog posts even though he is ranked low in terms of the number of blog posts, which reflects that the blog posts released by this user are recognized by users and the quality of the blog posts released is high, so they also have greater influence.

By comparing the research with the above three algorithms, it shows that the algorithm proposed in this paper is consistent with people's life perceptions, and can comprehensively evaluate the user influence from several factors, such as the direct influence of users, the influence indirectly generated by user fan relationship and the influence generated by blog post content, and improve the accuracy of user influence analysis in hot events.

5. CONCLUSION

Based on the hot events of Weibo, this paper proposes a microblog user influence analysis model that integrates user information and content information by integrating the users self-influence, the followers-influence, the content information and the sentiment polarity of comments. The experimental results show that the precision, recall and F1 value of UCRank have achieved the best results in the comparative experiments with multiple algorithms. It shows that the user influence analysis algorithm proposed in this paper can more improve the accuracy and effectiveness of user influence analysis in hot events by integrating multiple influence factors to comprehensively calculate user influence. In the future research work, we will conduct a more fine-grained analysis of user comments, so as to more accurately describe user emotions and more effectively monitor public opinion.

ACKNOWLEDGEMENTS

This work was supported by the National Social Science Fund of China [grant number 19BGL241]

REFERENCES

[1] Luarn, Pin , J. C. Yang , and Y. P. Chiu . "The network effect on information dissemination on social network sites." Computers in Human Behavior 37.37(2014):1–8.

[2] Li .Z, M. Li , and W. Ji . "Modelling the public opinion transmission on social networks under opinion leaders." Iop Conference 69(2017).

[3] Chen Z , Taylor K . Modeling the Spread of Influence for Independent Cascade Diffusion Process in Social Networks[C]// IEEE International Conference on Distributed Computing Systems Workshops. IEEE, 2017.

[4] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. 2010: 261-270.

[5] Shi lei, Zhang cong, Wei lin. Introducing Active Index to the Microbloggers' Ranking [J]. Journal of Chinese Computer Systems, 2012, 33(1):110-114.

[6] Kwak H, Lee C, Park H, et al. What Is Twitter, A Social Network or A News Media, 2017,31(04):184-190.

[7] Wu Hui, Zhang Shaowu, Lin Hongfei. Evaluation of the Users' Influence on Microblog [J]. journal of Chinese Information Processing,2017,31(04):184-190.

[8] Yang Tian, Zhao Limei. Research on the relationship between "double top" university library microblogs based on social network analysis [J]. Inner Mongolia Science and Technology and Economy, 2019, No.428 (10): 72-74

[9] Hao F, Chen M, Zhu C, et al. Discovering Influential Users in Micro-Blog Marketing with Influence Maximization Mechanism[C]. 2012 IEEE Global Communications Conference, 2012: 470-474.

[10] Pei, Li, Haichao, et al. Modeling and Estimating User Influence in Social Networks[J]. IEEE Access, 2020, 8:1-1.

[11] Jendoubi S, Martin A, Liétard L, et al. Two Evidential Data Based Models for Influence Maximization in Twitter [J]. Knowledge-Based Systems, 2017, 121(4):58-70.

[12] Wu H, Shang J, Zhou S, et al. IMPC: Influence Maximization Based on Multi-Neighbor Potential in Community Networks [J]. Physica A: Statistical Mechanics and Its Applications, 2018, 512(C):1085-1103.

[13] Simsek A, Kara R. Using Swarm Intelligence Algorithms to Detect Influential Individuals for Influence Maximization in Social Networks [J]. Expert Systems with Applications, 2018, 114(12):224-236.

[14] Boyd D, Golder S, Lotan G. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter[C]. 2010 43rd Hawaii International Conference on System Sciences, IEEE, 2010:1-10.

[15] Li Jun, Chen Zhen, Huang Jiwei Research on microblog influence evaluation [J] Information Network Security, 2012, (3): 10-13, 27

[16] Cano A E, Mazumdar S, Ciravegna F. Social Influence Analysis in Microblogging Platforms-A Topic-Sensitive Based Approach [J]. Semantic Web, 2014, 5(5):357-372.

[17] Rezaie B, Zahedi M, Mashayekhi H. Measuring Time-Sensitive User Influence in Twitter [J]. Knowledge and Information Systems, 2020, 62(9):3481-3508.

[18] Li S, Zhang Y, Jiang P, et al. Predicting User Influence in the Propagation of Toxic Information[C]. International Conference on Knowledge Science, Engineering and Management. Springer, Cham, 2020:459-470.

[19] Ouyang Chunping, Chen Xianglong, Liu Yongbin. Four-degree user influence analysis model based on online news review [J]. Computer Engineering and Design, 2021,42 (09): 2671-2678. DOI: 10.16208/j.issn1000-7024.2021.09.036

[20] Liu J, Dang Y, Wang Z, et al. Relationship between the in-degree and out-degree of WWW[J]. Physica A: Statistical Mechanics and its Applications, 2006, 371(2): 861-869.

[21] Weng J, Lim E P, Jiang J, et al. Twitterrank: finding topic-sensitive influential twitterers[C]//Proceedings of the third ACM international conference on Web search and data mining. 2010: 261-270.