IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 6, 2021, accepted April 20, 2021, date of publication April 26, 2021, date of current version May 13, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3075601

Tracking Attention of Social Media Event by Hidden Markov Model–Cases from Sina Weibo

YINGHONG MA, HUI JIAO, AND LE SONG
Business School, Shandong Normal University, Jinan 250014, China
Corresponding author: Yinghong Ma (yinghongma71@163.com)

This work was supported by the National Natural Science Foundation of China under Grant 71471106.

ABSTRACT Sina Weibo has significantly impact on the information diffusion processes in many real-world social events. A large number of active users on Sina Weibo not only push the opinion diffusion, but also increase the influence abilities of events which conversely attracted much attentions of users to follow them. How to effectively track the event attention of users is one of the most important channels to get the public opinions. In order to predict the event attention more accurately, motivated by observations of social events' influence concerning with users and microblogs, we quantify the user popularity from the four dimensions: the user activity, the user behavior, the user authenticity and the user infection ability. And the non-collinearity of these four dimensions is tested to ensure the comprehensiveness and non-redundancy of the evaluation. Then, combining with the logic framework of Hidden Markov Model, we propose an algorithm to predict the Weibo event attention by using the user popularity. Meanwhile, in order to better detect the performance of the prediction algorithm, we integrate the static and dynamic information of microblog content to directly quantify the current Weibo event attention as a benchmark, and the performance of four prediction algorithms (including our algorithm) is tested with six real data sets which are chosen from the popular events in China from 2019 to 2020. Through comparison, we find that the user popularity can be used to predict the event attention, and the Hidden Markov Model prediction method by using the user popularity shows good prediction performance.

INDEX TERMS User popularity, Weibo event attention, hidden Markov model.

I. INTRODUCTION

In recent years, with the rapid development of the Internet and the popularity of intelligent terminals, the Sina Weibo platform emerge in China as the times require. This platform not only provide a large number of new ways of information exchange for the public, but also become an important channel for people to obtain information [1]. However, the massive and rapid dissemination of information in Sina Weibo platform is a double-edged sword for the development of society. On the one hand, users can use the Weibo platform to spread the mainstream voice, which could promote the expression of public opinion and play a positive role [2]. On the other hand, the massive dissemination of false information in the Weibo platform will have a bad impact on society, and if such information is not controlled in time, it will cause a crisis [3]. In the face of this double-edged sword of social development, the Internet content governance has become an indispensable and effective means [4]. However, in the process of Internet content governance, it is impossible to control the content data of massive events one by one. In order to effectively achieve Internet content governance, it is necessary to use technical means to effectively obtain the attention of events, and adopt classified management for the contents that may become popular events. In the rapidly changing social media platform, it is very important for a country's stable development to accurately analyze which event will become a hot event in the large-scale event content data, and to track the future trend of the event's attention. For example, finding popular topics in real time and analyzing their public opinion tendency can help sociologists understand the social focus and people's mental state, so that the government can formulate corresponding policies to prevent and control malignant events [5]; it is of great practical significance to predict and track the attention of public emergencies in time for the protection of national political, economic and social security [6].

Because of the importance of tracking event attention, many scholars have conducted in-depth research on it. Based on the study of event types, many scholars have provided rich

The associate editor coordinating the review of this manuscript and approving it for publication was Fabrizio Messina.

methods for quantifying and predicting attention trends of Weibo events. Cao proposed a method to divide messages into different groups, and then trained a group-specific popularity prediction model for each group of messages [7]. Liu took an improved state transition-based algorithm to predict the popularity of different types of events [8]. However, after classifying events, different model parameters are set for different event categories, which greatly reduces the universality of the model. If the classification of events is wrong, its prediction ability will be affected.

Therefore, the method based on the microblog content is more widely used. The related characteristics of microblog content can be summarized as static characteristics and dynamic characteristics. Some previous analysis on the attention of Weibo events are concerned on the static characteristics of microblogs, such as the number of out-links, the length of blog posts, the content and the topical information of messages [9]. For example, Cui analyzed the number of hashtag (topic keywords beginning with #) in microblogs on Sina Weibo platform to study the popularity of popular event keywords discussed by users, so as to predict the trend of public attention to the related event [10]. While the real-world events are dynamics and diversify, the statics features are inevitable got into the bias of predication. Thus, the dynamic characteristics of microblogs, such as the number of retweets and pageviews, are used to measure the degree of user participation in information dissemination [11]. For example, Lu took the microblog content as the research object, and analyzed several new implicit factors that affect the popularity of content, such as modality, maximum media weight and activeness, then modelled the reporting process in a time dynamic way to realize the prediction of event attention in Sina Weibo [12]. Using the microblog content to predict the event attention can improve the universality of prediction methods or models. However, microblog contents are released by users, and whether the final event can be concerned or not, users play a decisive role. Therefore, only relying on the microblog content to build a prediction model is one-sided. It is more comprehensive to predict the evolution trend of event attention by combining the information of users and microblogs.

Media users are makers and disseminators of Weibo events, the increment of Weibo event attention largely depends on the popularity and participation of users themselves. Weibo events were attracted by a large number of users and diffused by them, and conversely, user popularity enhanced the attention of event. Therefore, there is a natural connection between user's popularity and Weibo event's attention. However, few of the previous researches were related the relationship of them. Therefore, the quantification of user popularity is very important, which have been studied in many ways. Such as the Twitter users were measured by the properties of their posts, retweets and the number of mentions [13], or by the retweet relationship and the retweet distance [14]. The time for the released information of event, the number of users' followers, the influence of followers, and the degree of attention of followers were used to measure the user popularity [15]. However, most of the existing researches on the quantification of user popularity are based on user attributes or text content attributes, and few researches can integrate these two dimensions of attributes. In addition, in the existing user popularity evaluation indicators, the parameter setting is often only concerned with the selected attribute dimension, and ignores the collinearity between the parameters.

Inspired by the above review, how to integrate user attributes and content attributes to achieve a comprehensive quantification of user popularity, how to avoid the redundancy of multi parameter evaluation in the process of integration, and how to use user popularity to achieve the tracking of event attention are the key issues of this paper. Therefore, from the perspective of users, we integrate user attributes and content attributes to quantify the user popularity. In the process of quantifying user popularity, we design four calculation parameters and make non-collinear test on them to ensure their independence and uniqueness. These non-collinear parameters are used to construct a non-linear user popularity metric, which is used as a dominant variable to predict Weibo event attention combined with the logical framework of Hidden Markov Model (HMM).

Three contributions are presented in this work: (1) we present a user popularity metric to measure user influence by combining comprehensively the characteristics of microblogs and users; (2) the Weibo event attention degree is defined on original microblogs and retweet microblogs; (3) in the framework of HMM, the original observation sequence (user popularity) is used to predict the hidden state sequence (Weibo event attention).

The rest of this paper is arranged as follows: in Section II introduce the research data sets including attributes and statistical characteristics of the data; in Section III, two measurements, the user popularity and the event attention, are deliberated constructed based the feathers of users and microblogs; in Section IV, we present an algorithm based on HMM to predicate the Weibo event attention trends with the observations of user popularity. Section V, the feasibility of the event attention predication algorithm is simulated by 6 cases collected in this work. And in last section VI, we compare our algorithm with other 3 methods which proved the accuracy of it. Conclusions and limitations of this work are also considered in the final section.

II. CASES FROM WEIBO MEDIA

Six different categories of data sets covering social managements, entertainments, disasters, educations, public healths and managements, collected from the Sina Weibo are used for the research, and the time span of the data is one month, shown in Table 9 of Appendix 1.

A. DATA COLLECTION

The data collected in this paper are six different categories of popular events in Sina Weibo, and the six events are as follows: ''Xi'an Benz Rights Protection'', ''Tianlin Zhai

academic fraud'', ''African locust plague'', ''2020 college entrance examination postponed'', ''Fang Fang's Diary'' and ''Teacher salary is not lower than civil servant''. The data were all taken within a month after the event was exposed [16]. The data generation time, and the amount of data of specific event are shown in Table 9 of Appendix 1. In order to carry out the follow-up research more accurately, we divided microblogs into original microblogs and retweet microblogs. The original microblog is a microblog that users originally published after their own thinking and compilation; the retweet microblog is a microblog that users directly retweet to microblogs of others without compiling by themselves. There are both similarities and differences in their attributes. For an original microblog, its attributes mainly include the content of a microblog, the release time of a microblog, the number of retweets, the number of comments and the number of likes of a microblog. For a retweet microblog, not only the above attributes must be considered, but also the time difference with the original microblog need be considered. For the attributes of Weibo users, this article analyzes the user's name, the number of user's followers, the number of user's following, the number of microblogs posted by the user, and whether the user is an authenticated user.

B. STATISTICAL PROPERTIES

In this part, we analyze the statistical characteristics of microblogs and users. In the statistical analysis of microblogs, we analyze the number of original microblogs and retweet microblogs contained in each event, and the number of topics, the number of @, the number of out-links of different types of microblog for different events. In the statistical analysis of users, we analyzed the number of authenticated users and non-authenticated users, and the average number of fans for different types of users in different events. According to the data captured above, the statistical characteristics of different Weibo events are shown in Table 10 of Appendix 1.

III. EVALUATIONS ON EVENTS

The user popularity and the Weibo event attention are introduced to measure the impact of users and events respectively. The feasibility of the variables are also verified by cases from Weibo media, of which notations and definitions are shown in Table 1.

A. USER POPULARITY

Many scholars have put forward their own understanding of the user popularity. Wei *et al*. divided the user popularity into the influence of followers, comments, and retweets [15]. Francalanci *et al*. thought that the user popularity refers to the extent to which users share text, pictures and other content on other users [17]. Here, we suppose that the user popularity is the influence of users in social media, which involved by their activity, behavior information, authenticity, and infection ability.

TABLE 1. Notations and definitions.

Notation	Definition, where ♯ is the short of "the number of".
y_t, z_t	♯ original microblogs, retweets;
T	the length of lasting time of the microblog event;
x_{iz}, x_{ip}, x_{id}	♯ retweets, comments, likes of microblog i;
$aw_u, fans_u$	♯ posted microblogs, followers of user u;
N_{theme}	♯ topics mentioned,
$N_{mention}$	♯ of @ in a microblog;
N_{link}	♯ out-links involved a microblog;
T_v	the release time of a microblog v;
T_{fw}	the release time of the first microblog;
LEN_v, L	♯ words of microblog v, the maximum length;
ω	ω=LEN_v/L is the content fullness of microblog v;
T_i	The release times of the retweet microblog,
T_I	The release time of roriginal microblog to retweet;
m, m_r	♯ original microblogs, retweet microblogs;
n, N	♯ events related to the posted microblog, total per hour.

(1) User's activity A_u. The more microblogs users post, the bigger influence they have on others [18]. Therefore, user's activity could be measured by the number of original microblogs and the retweet microblogs related to an event. Then, A_u is defined by equation (1),

$$A_u = \frac{\sum_{t\in T} y_t + \sum_{t\in T} z_t}{T}. \tag{1}$$

(2) User's behavior information P_u. User's behaviors include the retweet, comment and like. Retweeting and commenting indicate that other users participate in discussion on the event, or liking indicates that the event appeals to them or simply has been seen. As a result, the retweet, comment and like are regarded as behaviors of a user. Studies show that the three behaviors have different effects for users. The total number of retweets of a microblog is one of the most special indicators of the user popularity [19]. P_u is defined by the formula (2),

$$P_u = \frac{\omega_1 \sum_{i=1}^{n} x_{iz} + \omega_2 \sum_{i=1}^{n} x_{ip} + \omega_3 \sum_{i=1}^{n} x_{id}}{n}, \tag{2}$$

where $\omega_1, \omega_2, \omega_3$ are three adjust indicators, and $\omega_1 + \omega_2 + \omega_3 = 1$.

(3) User's authenticity V_u. Some scholars proved that the number of original microblogs of a user enhances his/her authenticity score [20]. And on the other hand, the greater influence of a authenticated user will display much more desire in discussion of events [21].

In order to improve the authenticity and reliability of the microblog content, Sina Weibo has launched an authentication system, and parted users into two classes: authenticated users and the unauthenticated respectively. Therefore, we measure user's authenticity respecting these two classes:

$$V_u = \begin{cases} a \cdot \dfrac{\sum_{t\in T} y_t}{\sum_{t\in T} y_t + \sum_{t\in T} z_t}, & \text{authenticated user;} \\ b \cdot \dfrac{\sum_{t\in T} y_t}{\sum_{t\in T} y_t + \sum_{t\in T} z_t}, & \text{unauthenticated user,} \end{cases} \tag{3}$$

where a, b are the ratios of authenticated users and the unauthenticated users, respectively, $a + b = 1$ [6].

(4) User's infectious ability G_u. In social media, users influence their followers by their posted or retweeted microblogs, and attract them to participate the event discussion. Hence, the number of followers is regarded as one of the most intuitive measurements for the user popularity [18]. And, the number of posted microblogs by users also is a measurement to show the infectious ability of the user. To sum up, we use the number of followers $fans_u$ and the total number of posted microblogs aw_u to qualify user's infectious ability, shown in formula (4),

$$G_u = fans_u + aw_u. \tag{4}$$

The user's activity, behavior information, authenticity and infectious ability are defined in different dimensions and orders. On the one hand, to avoid the incomparable, we adjust them by the min-max normalization method [22] such that the four indexes fall into the interval [0, 1], and are still denoted by A_u, P_u, V_u, G_u. On the other hand, to avoid the redundancy, we compute Pearson correlations of them shown in Table 2 by the data sets of Weibo users and events appended in Table 10 of Appendix 1.

TABLE 2. Pearson correlation of the four indexes.

	A_u	P_u	G_u	V_u
A_u	1.000	0.040	0.128	0.213
P_u	0.040	1.000	0.215	0.036
G_u	0.106	0.192	1.000	0.175
V_u	0.213	0.036	0.126	1.000

The largest value of the correlation between variables is 0.215, P_u and G_u, which implies the linear correlation is week. In other words, no index can be replaced by another. Therefore, user popularity Y_u is measured by the four indexes comprehensively with formula (5).

$$Y_u = V_u(\gamma_1 A_u + \gamma_2 P_u + \gamma_3 G_u), \tag{5}$$

where $\gamma_1 + \gamma_2 + \gamma_3 = 1$ and $\gamma_j \geq 0$ for $j = 1, 2, 3$.

Again, we use the data sets of Weibo users and events appended in Table 10 to test the fitness of the Y_u by the ordinary least squares fitting process [23]. The R-squared of the Y_u model is 0.9844 and the p-values of A_u, P_u, G_u and V_u are 0.00, 0.00, 0.00 and 0.013, respectively. Hence, Y_u comprehensively expressed by the four indexes is acceptable.

In formulas (2) and (5), ω_i and γ_j are adjust coefficients in computing Y_u, $i, j = 1, 2, 3$. Based on the real data of users, the values of them can be obtained by Entropy Weight Method. The specific steps of Entropy Weight Method can be found in [24], which can reduce the subjectivity and randomness for weighting the indexes' weighs.

B. THE EVENT ATTENTION DEGREE

The event was something that occurred at a specific time, place, and could be attracted by media attention, or might be widely discussed on social media [25]. Besides those characters, we also suppose that the event is involved by certain objects and diffusion quickly online. The event attention is the degree of events to be discussed, which can be presented by the number of microblogs related to the event, by static metrics related to the content of microblogs, and by dynamic metrics corresponded to the information of retweets and comments. In this subsection, we define Weibo event attention by using the static and dynamic metrics of events.

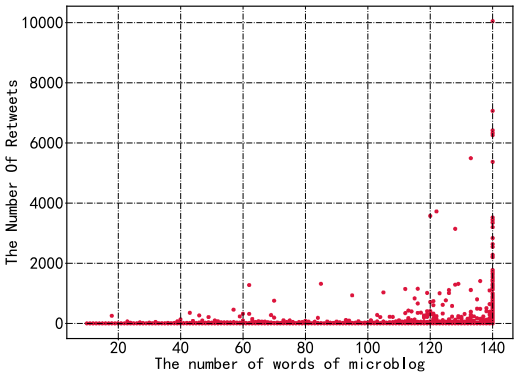
1) THE STATIC AND DYNAMIC METRICS

Some scholars had shown that the number of themes and mentioned peoples have positive impacts on the event [26], while out-links reduces these impacts [27]. Agarwal showed that words and comments of the influential microblog have more lasting impact than other microblogs [28]. When a microblog was published, the number of themes N_{theme}, the number of mentioned peoples $N_{mention}$, the number of involved out-links N_{link}, and the number of words LEN_v of the microblog are all fixed, which represent a certain of influences of Weibo event. Hence, the static metrics represent some certain influence of the Weibo event.

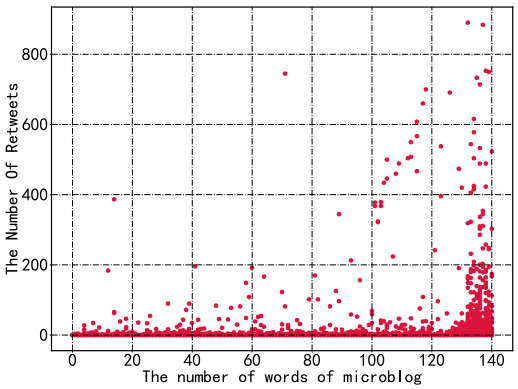
Therefore, we use the data set to analyze the relationship between the length of the microblog and the amount of retweets x_{iz}, shown in Figure 1. Most of retweets are concentrated in the range of 0 to 1000, and only 1.104% of them are greater than 1000. The case of microblogs with retweets' number ranging from 0 to 1000, it shows a positive correlation between the number of retweets and the length of the microblog. The more words there in the microblog text are, the more likely microblogs to be retweeted, shown in Figure 1(b). For the microblogs with the number of retweets less than 200 shown in Figure 1(c), are similar with the case in Figure 1(b).

On the Sina Weibo, the number of new microblogs begin to rise significantly from 8:00 to 9:00, and since then, fluctuates at a high level. It was shown that the greater of the number of retweets x_{iz} and the number of comments x_{ip}, the greater the attention of Weibo events [15]. It seemed that Weibo event attention is affected by the release time of the related microblogs. In order to verify the observation, the percentage of the number of published microblogs in each hour to the total number in a day is used to simulate the distributions of microblogs and Weibo posts. Figure 2 shows the statistical results based on the data sets of Table 9 in Appendix 1. Results in Figure 2(a) show that the number of microblogs posted by users from 8:00 to 22:00 is relatively high, which might enhance the possibility of users' interaction. Therefore, the release time T_v of microblogs affect the attention of Weibo events.

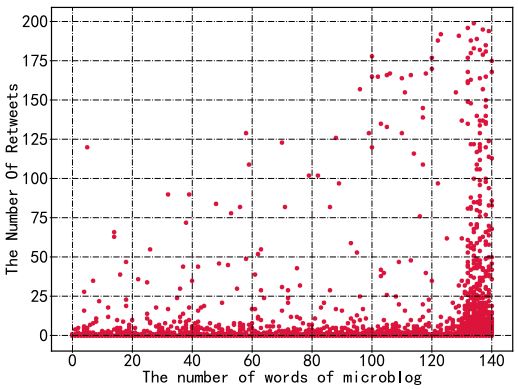
The information of microblogs has strong timeliness and the content of information loses fresh over time [29], which makes difficult for people to continuously pay their attention on a given microblog. To test this phenomenon, according the classification of users in [30], the statistics of the number of retweets for a single microblog about eight classes of users at different times are shown in Figure 2(b). Distributions of the microblog retweets show a sharp decrease with the time going and tends to zero after published about 7 hours in a



(a) Distribution of microblogs and their retweets.



(b) Distribution of retweets less than 1000.



(c) Distribution of retweets less than 200.

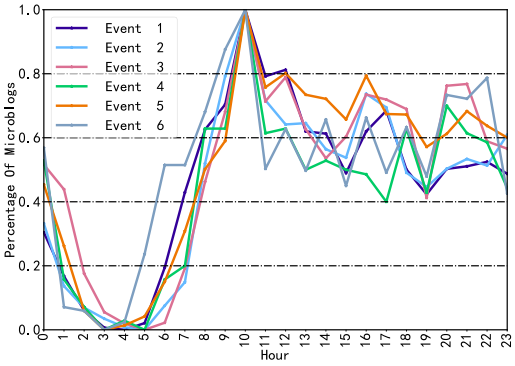
FIGURE 1. Distributions with different scales of the length of microblogs and the number of retweets.

day, which implies that microblogs have strong timeliness in different time of a day.

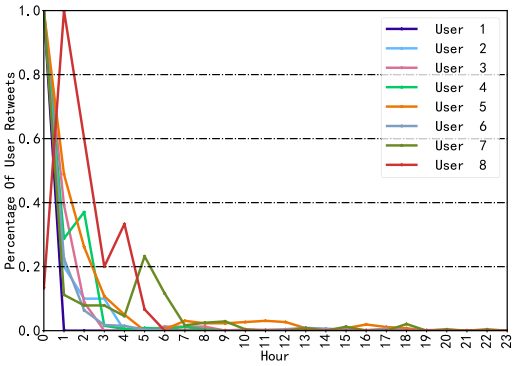
2) THE ATTENTION DEGREES OF ORIGINAL MICROBLOG AND RETWEET MICROBLOG

We define the attention of original microblog respecting to static and dynamic metrics, named microblog static attention and dynamic attention, denoted by *WSA* and *WDA* in equations (6) and (7), respectively.

$$WSA = \alpha_1 N_{theme} + \alpha_2 N_{mention} - \alpha_3 N_{link}, \tag{6}$$



(a) Distribution in events.



(b) Distribution in users.

FIGURE 2. (a) is the distribution of microblog posts in different events, and the data set of the total number of microblog posted per hour of the 6 Weibo events are shown in of appendix 1. (b) is the distribution of microblogs retweets in different classes's users, and all users are parted into 4 classes by their sizes and parted into 8 classes by their user-authentication, shown in table of appendix 1.

$$WDA_{om} = (\beta_1 x_{iz} + \beta_2 x_{ip}) \cdot \frac{N(t)}{\sum_{t=1}^{T} N(t)} \cdot \frac{1}{T_v - T_{fv} + 1}, \tag{7}$$

where α_1, α_2, α_3, β_1 and β_2 are adjustment coefficients in computing *WSA* and *WDA*, which are obtained by using the Entropy Weight Method based on the real data about microblogs in the same way as ω_i and γ_j.

Therefore, the attention degree of an original microblog is defined by two sides *WSA* and WDA_{om}, denoted by YA_{om} shown as formula (8).

$$YA_{om}(i) = \begin{cases} \omega(WSA + WDA_{om}), & x_{iz} \ or \ x_{ip} \neq 0, \\ 0 & x_{iz} \ and \ x_{ip} = 0, \end{cases} \tag{8}$$

where ω is the content fullness of a microblog defined in Table 1. In formula (8), the Weibo attention is interpreted by the retweets and comments, which can significantly promote the attention of events; otherwise, the microblog is ignored or no influence on events. Because Weibo platform does not show the reading times of microblogs, so the user interaction usually is measured by retweets [15].

Some basic attributes of retweet microblogs are the same as the original microblog, but the release time of the retweet and original microblog is different. When the release time difference between the original microblog and the retweet

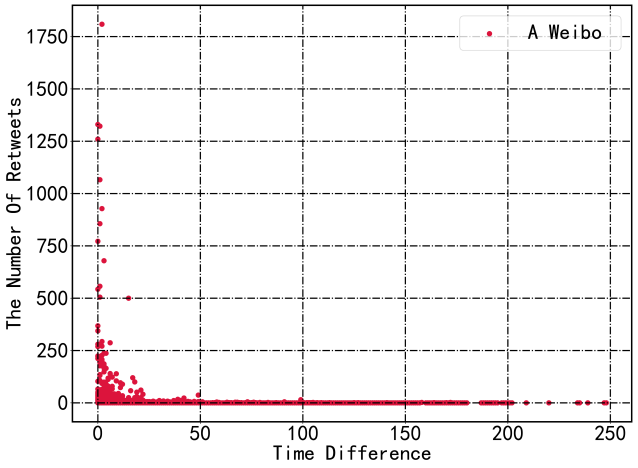


FIGURE 3. The relationship between the difference of release time and the number of retweets for all time.

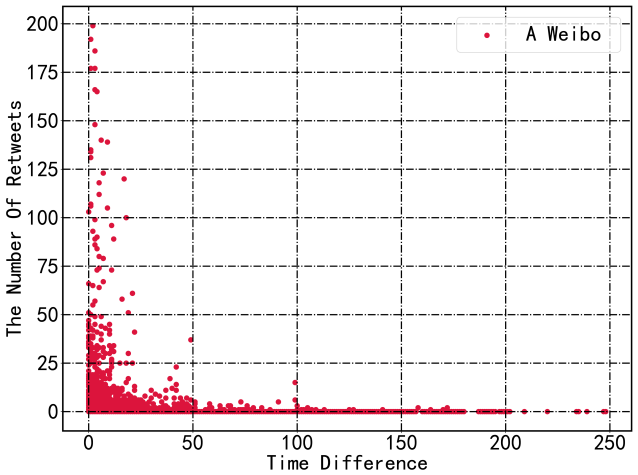


FIGURE 4. This figure shows the case of number of retweets less than 200 of figure 3.

microblog is small, retweet microblogs would have a high attention. Then, the retweet microblog is more likely to be followed by other users. So, we analyze the relationship between the release time difference and the number of retweets, shown in Figure 3. Most of microblogs are retweeted no more than 200 times, and only 0.37% of them are retweeted greater than 200. Figure 4 shows some relations between the number of retweets and the release time difference, but most of the microblog retweets are in the lower part. Meanwhile, when the number of retweets less than 20 is noticed, it is can be found that the greater the difference of publishing time, the smaller the number of microblog retweets, which is shown in Figure 5.

Inspired by the analysis on the difference of release time and the number of retweets, we find the retweet microblog differs from the original microblog only in the dynamic factors, such as the release time. Then, we define the dynamic attention degree of the retweet microblog as formula (9).

$$WDA_{rm} = (\beta_1 x_{iz} + \beta_2 x_{ip}) \cdot \frac{N(t)}{\sum_{t=1}^{T} N(t)} \cdot \frac{1}{T_i - T_I + 1}. \quad (9)$$

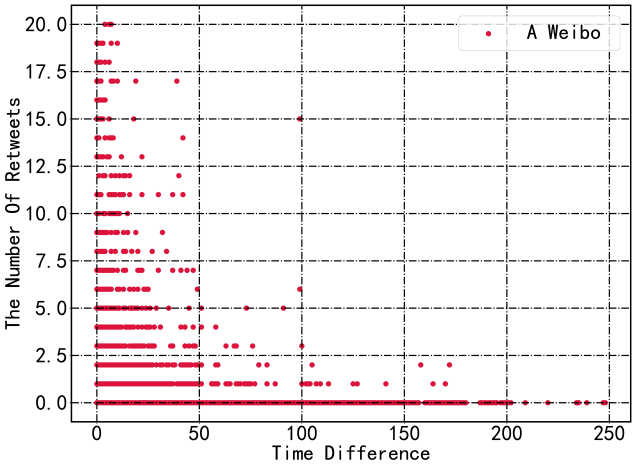


FIGURE 5. It shows the case of retweets less than 20 of figure 3.

Combing the static and the dynamic attention degree of retweet microblogs, the attention degree of a retweet microblog is defined as

$$YA_{rm}(i_r) = \begin{cases} \omega(WSA + WDA_{rm}) & x_{iz} \ or \ x_{ip} \neq 0, \\ 0 & x_{iz} \ and \ x_{ip} = 0. \end{cases} \quad (10)$$

3) THE DEFINITION ON EVENT ATTENTION DEGREE

When a social event is wildly spread by users through social media, enormous discussions and opinions would emergence. Different opinions represent different sides of social cognitions for the event, and integrate different topics. As a result, in order to increase their exposure, some users even use the popular topics to advertise maliciously. However, the contents of such microblogs are irrelevant to the event, nor contribution to the event attention degree. Therefore, the topic contained in microblog need to be identified, and the microblog unrelated to the event need to be eliminated when measuring event attention. We cluster microblogs into different topics based their correlation with the event, and revise the formula (8). Topics are detected in two steps: we segment the text content of microblogs by "jieba" library of python, and delete the stop words to get the word set; then we cluster the word set into several categories by k-means method [31], which is regarded as the topics of the microblogs. In particular, the number of the topics are determined by SSE (sum of the squared errors) [32] and silhouette coefficient [33].

In general, the content of the retweet microblog is almost as same as the original microblog. This leads to the overlap measurements of YA_{om} and YA_{rm}, which requires to revise the definition of the attention of the original microblog i, $YA_{om}(i)$. When we calculate the impact of microblog on event attention, we need to revise the measurement of the attention of the original microblog $YA_{om}(i)$. So we take the average microblog similarity on original microblog i into consideration to revise the formula of microblog's attention. Then revised attention

of a single original microblog i is shown in formula (11).

$$YA'_{om}(i) \leftarrow YA_{om}(i)\frac{1}{p}\sum_{j=1,j\neq i}^{p-1} S(d_i, d_j), \tag{11}$$

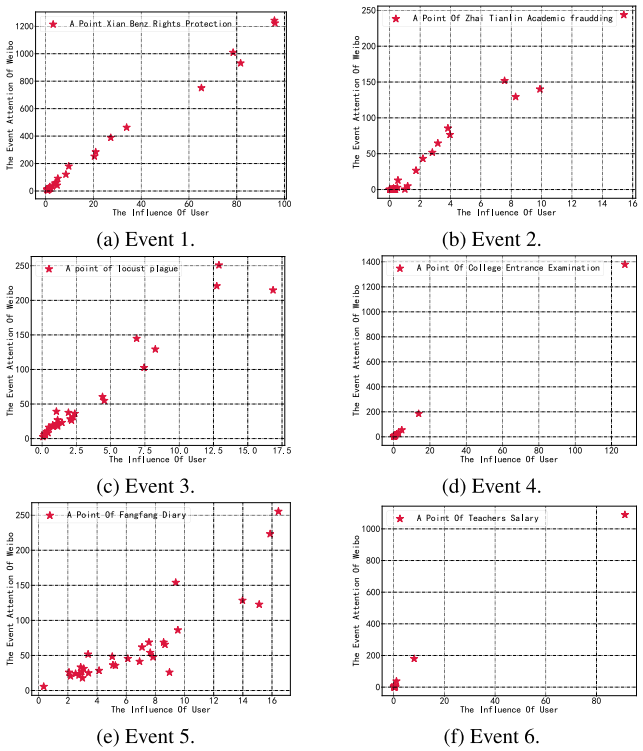
where p is the number of microblogs on topic i, $S(d_i, d_j)$ is the cosine similarity of microblog j and topic i.

Finally, we define the event attention by combining the original microblog attention and the retweet microblog attention and get the formula (12):

$$EA = \sum_{i=1}^{m} YA'_{om}(i) + \sum_{i_r=1}^{m_r} YA_{rm}(i_r). \tag{12}$$

C. RELATIONS BETWEEN USER POPULARITY AND EVENT ATTENTION

In order to further explore the relationship between the user popularity and event attention, we calculate the user popularity and event attention of the six Weibo events, and plot scatter diagrams corresponding to the two variables, shown in Figure 6, which display correlation relationships of them.



(a) Event 1. (b) Event 2.

(c) Event 3. (d) Event 4.

(e) Event 5. (f) Event 6.

FIGURE 6. The relationships between the user popularity and the attention of Weibo events.

The correlation of the Weibo event attention and user popularity is modeled as a polynomial regression equation, shown in Table 15 of Appendix 1. With the scatter figures and the polynomial regression equations, we declare that the user popularity and the event attention have a certain correlation. Thus, it is possible to track Weibo event attention by user popularity.

IV. ALGORITHM OF TRACKING ATTENTION WITH HMM

As we all know, all kinds of events in Sina Weibo platform are affected by various exogenous factors, such as the influence of other events, the change of topic types, the psychological role of users and so on. Because the attention of Weibo events shows variability, it is unable to achieve accurate long-term prediction of event attention. Therefore, the short-term prediction method need to be used to estimate the attention of Weibo events. Hidden Markov Model (simply HMM) with good short-term prediction effect can provide a logical framework for predicting Weibo event attention.

The user popularity and event attention are two sides for the event prediction. The former is a relatively fixed matric, and the latter cannot be determined alone because of the diversity of retweets of microblogs or the other involving topics. That naturally emerges two questions: how the event evolves or what time we should take strategies to control the event spreading. Hence, we predict the event evolving or the control time by HMM which depends on the observation sequence of the user popularity statuses to predict the hidden sequence of the event short-term evolving trend.

HMM is a two-level stochastic process, one is the observed state sequence and the other is the hidden or the forecast state sequence, denoted by $O(t)$ and $S(t)$ respectively, to mine or predict information according the obtained data from the system [34]. A visualized example of HMM is shown in Figure 7.

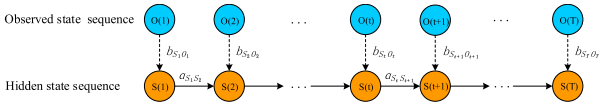


FIGURE 7. Hidden Markov Model (S, O, A, B, π), where $S = (S_1, S_2, \cdots, S_c)$, $O = (O_1, O_2, \cdots, O_c)$ are sequences of the hidden state and the observation state, respectively. $A = (a_{S_tS_{t+1}})$ is the state transition probability matrix, where $a_{S_tS_{t+1}}$ is the transition probability from observation states S_t to S_{t+1}. $B = (b_{S_tO_t})$ is the output probability from an observation state O_t to a hidden state S_t. $\pi = \{\pi_1, \cdots, \pi_i, \cdots, \}$ is the initial probability vector, and π_i is the probability of the state s_i coming out at time t.

For example, in this paper, we define three states of the event attention and user popularity at a certain time t: up, down and unchange respecting to the values is going up, going down and keep unchange comparing with the values at time $t-1$. Then $c = 3$, $O = \{o_1 = up, o_2 = unchange, o_3 = down\}$ and $S = \{s_1 = up, s_2 = unchange, s_3 = down\}$. We encode the states' values of user popularity and the event attention: Let $O_{up} = 2$ when $Y_u(t) > Y_u(t-1)$; $O_{unchange} = 1$ when $Y_u(t) = Y_u(t-1)$; $O_{down} = 0$ when $Y_u(t) < Y_u(t-1)$. And, $S_{up} = 2$ when $EA(t) > EA(t-1)$; $S_{unchange} = 1$ when $EA(t) = EA(t-1)$; $S_{down} = 0$ when $EA(t) < EA(t-1)$. Then state transition probability matrix $A = (a_{ij})_{3\times3} = (P(S_t = s_j|S_{t-1} = s_i))_{3\times3}$, where $s_i, s_j \in S$ and $\sum_j^3 a_{ij} = 1$ for $i = 1, 2, 3$. $B = (b_{ij})_{3\times3} = (P(O_t = o_j|S_t = s_i))_{c\times c}$ is the emission probability matrix from an observation state to the hidden state. Denote $P(S_t = s_j|S_{t-1} = s_i)$ and $P(O_t = o_j|S_t = s_i)$ by $P(s_i \to s_j)$ and $P(o_j|s_i)$ respectively. According

to the above definitions and analysis, A and B are the 3×3 transition probability matrixes shown as follows:

$$A = \begin{pmatrix} P(s_3 \to s_3)\ P(s_3 \to s_2)\ P(s_3 \to s_1) \\ P(s_2 \to s_3)\ P(s_2 \to s_2)\ P(s_2 \to s_1) \\ P(s_1 \to s_3)\ P(s_1 \to s_2)\ P(s_1 \to s_1) \end{pmatrix},$$
$$B = \begin{pmatrix} P(o_3|s_3)\ P(o_3|s_2)\ P(o_3|s_1) \\ P(o_2|s_3)\ P(o_2|s_2)\ P(o_2|s_1) \\ P(o_1|s_3)\ P(o_1|s_2)\ P(o_1|s_1) \end{pmatrix}. \quad (13)$$

By the above notations and definitions, we design an algorithm to calculate values and trends of the event attention by Baum-Welch algorithm on HMM process [35], shown from **step 1** to **step 6** in Table 3.

The complexity of this algorithm includes the time complexity and space complexity. The time cost is $O(n_u)$ in calculating the metrics of user popularity, A_u, P_u, V_u and G_u; the space complexity is $O(n_u)$ too. The time and space complexities to calculate Weibo event attention are $O({n_m}^2)$, where n_u and n_m are the number of users and microblogs, respectively.

Since Baum-Welch algorithm on HMM process is a recursive algorithm, and the probability at $t+1$ time calculation only needs the results at time t. Suppose that there are c hidden states $\{s_1, s_2 \cdots, s_c\}$ and T observations $\{O_1, O_2, \cdots, O_T\}$. Then the complexity of calculating $\lambda(A, B, \pi)$ during the process of HMM is $O(Tc^2)$. Finally, the complexity of the algorithm is $O(Tc^2)+2O({n_m}^2)+O(n_u)$.

V. EXPERIMENTS RESULTS

A. TREND OF WEIBO EVENTS ATTENTION

In the above Weibo event data that was introduced in Section II-A, we divide the data set of each Weibo event into 5:1 in chronological order, we use the first 5/6 time series data as the training set to training the parameters $\lambda = (A, B, \pi)$ of the HMM and the last 1/6 time series data as the test set to test the accuracy of the prediction. According to formula (5), the user popularity state sequence of each event can be counted, where 0 means down, 1 means unchange, 2 means up, as shown in Table 4. The user popularity state sequence is used as the observation variable, and the initial parameters $\lambda_0 = (A_0, B_0, \pi_0)$ are input into the Baum-Welch algorithm to learn the parameters $\lambda = (A, B, \pi)$ of the HMM, and finally the Hidden Markov Model of each Weibo event is obtained.

Taking Event 1 "Xi'an Benz Rights Protection" as an example, its the state transition matrix of Weibo event attention is shown in table 5 and user popularity output probability matrix is shown in Table 6, and the state transition matrix of other event attention and the output probability matrix of user popularity are shown in Table 12 and 13 of Appendix 1.

After the HMM parameters $\lambda = (A, B, \pi)$ are determined, the event attention state prediction is performed for the remaining 20% of the time period. Same as above, 0 means down, 1 means unchange, and 2 means up. The forecast results are shown in Table 7 and Figure 8.

TABLE 3. HMM algorithm: predicating Weibo event attention.

Input: Contents of the event microblogs, the release time, retweets' numbers, comments or likes and the time difference with the given original microblog for retweets, Weibo users' name, followers, microblog retweeted by the user, and authenticated or not. Denote $O = \{up, steady, down\}$ be three states of the user popularity at each time, and the initial state HMM $\lambda_0 = (A_0, B_0, \pi_0)$.
Output: The trend predication of Weibo event attention, $\lambda = (A, B, \pi)$.

step 1: Calculate the event microblogs and users' popularity, $y_t, z_t, x_{iz}, x_{ip}, x_{id}, aw_u, fans_u$.

step 2: Denote $O = \{O_1, \cdots, O_t, \cdots, O_T\}$ be observing sequence. Let $\alpha_i(t)$ be a forward probability at the first t observation, and the $t+1$ length sequences be $\alpha_j(t+1)$. At time t, we denote the probability of Markov chain state S_i in S by $P(O_1 O_2 \cdots O_t S_t|\lambda)$. Then, their formulas are shown in equation (14).

$$\begin{aligned} \alpha_i(t=1) &= \pi_i b_i(o_1), \\ \alpha_j(t+1) &= \sum_{i=1}^{T} a_i(t) a_{ij} b_j(O_{t+1}), \\ P(O|\lambda) &= \sum_{i=1}^{c} \alpha_i(T). \end{aligned} \quad (14)$$

step 3: Let $\beta_i(t)$ be the backward probability of the state i being in the remaining observation sequence $O = \{O_{t+1}, \cdots, O_T\}$, and $\beta_i(T)$ be the initial, respectively.

$$\begin{aligned} \beta_i(T) &= 1, \\ \beta_i(t) &= \sum_{j=1}^{c} a_{ij} \beta_j(t+1) b_j(O_{t+1}). \end{aligned} \quad (15)$$

step 4: We define the state transition from i to j as $\zeta_{i\to j}(t)$ and the probability of the state i at time t is $\varphi_i(t)$ in the state sequence. Then,

$$\begin{aligned} \zeta_{i\to j}(t) &= \frac{\alpha_i(t) a_{i,j} b_j(O_{t+1}) \beta_j(t+1)}{\sum_{i=1}^{c} \sum_{j=1}^{c} \alpha_i(t) a_{i,j} b_j(O_{t+1}) \beta_j(t+1)}, \\ \varphi_i(t) &= \sum_{j=1}^{c} \zeta_{i\to j}(t). \end{aligned} \quad (16)$$

step 5: Then an estimation for Weibo event attention $\widehat{\lambda} = (\widehat{A}, \widehat{B}, \widehat{\pi})$ in HMM could be updated, where elements of $\widehat{A}, \widehat{B}, \widehat{\pi}$ are obtained by equations (17), respectively.

$$\begin{aligned} \widehat{a}_{ij} &= \frac{\sum_{t=1}^{T} \zeta_{i\to j}(t)}{\sum_{t=1}^{T} \varphi_i(t)}, \\ \widehat{b}_j &= \frac{\sum_{t=1, O_t = O_j}^{T} \varphi_j(t)}{\sum_{t=1}^{T} \varphi_j(t)}, \\ \widehat{\pi}_i &= \varphi_i(1). \end{aligned} \quad (17)$$

step 6: Calculate the difference between the expected $P(O|\lambda)|$ and the estimated probability $P(O|\widehat{\lambda})$. If $|P(O|\widehat{\lambda}) - P(O|\lambda)| \le \epsilon$ for any given positive value of ϵ, output $\widehat{\lambda} = (\widehat{A}, \widehat{B}, \widehat{\pi})$ be the Weibo event attention predication; otherwise, set $i = i+1$ and go to **step 2**.

B. EVALUATION OF THE RESULTS

1) CREDIBILITY TEST OF HMM ALGORITHM

The Viterbi algorithm [36] can be used to evaluate the credibility of the HMM prediction algorithm. We input HMM parameter $\lambda(A, B, \pi)$ and the user popularity sequence in the training set, and with Viterbi algorithm, the event attention

TABLE 4. User popularity status sequence of different Weibo events.

Events	User popularity status sequence
Event 1	22222020020000200012020220
Event 2	20220220000020220102020020
Event 3	22200001020000220002000020
Event 4	20000222020022002000222201
Event 5	22000020012022000000220200
Event 6	20000020020122020202020022

TABLE 5. Attention state transition probability matrix *A* of event1.

States of EA in S	up	unchange	down
up	0.45439	0.09105	0.45456
unchange	0.00033	0.00033	0.99934
down	0.58319	0.00034	0.41647

TABLE 6. The probability matrix *B* of emitting the visible state of event1.

States of Y_u in O	up of EA	unchange of EA	down of EA
up of Y_u	0.99867	0.00041	0.00092
unchange of Y_u	0.00089	0.99853	0.00058
down of Y_u	0.00083	0.00041	0.99876

sequence can be output, the more the number of the same sequences, the higher the credibility of the HMM algorithm. First, by formula (5), the user popularity state sequence can be calculated. Second, The user popularity state is input to the Viterbi algorithm together with HMM parameter $\lambda(A, B, \pi)$, and the Weibo event attention state sequence with the highest probability is output. The real event attention implied in the training set is calculated by formula (12). Then, the calculated event attention state sequence is compared with the real event attention. The test results are shown in the Table 8. After that, the credibility of our algorithm to track the attention of six events is above 80%. Therefore, our HMM algorithm can be used to track the development of Weibo event attention, and its prediction results are credible.

2) ACCURACY ANALYSIS OF HMM ALGORITHM

In Section III, we defined the degree of event attention and quantitatively expressed it with the formula (12). The calculated value can be regarded as the real values of event attention, and the development trend of event can be obtained by using the change of the real value. According to the comparison between the HMM prediction sequence and the actual sequence of the attention of Weibo events, as shown in Table 7, it can be seen that the prediction results are stable and accurate.

What's more, our algorithm shows better short-term prediction ability, even using different sizes of training sets, our algorithm still has good performance. We provide data analysis support for the above conclusion in Appendix 2.

VI. DISCUSSIONS AND CONCLUSION

In previous Weibo event attention prediction, the mainly methods include grey model [37], neural network [38], ARIMA model [39], etc. In order to verify the superiority

TABLE 7. Comparing HMM algorithm with grey, ARIMA and neural network model.

Events	Actual	Grey	ARIMA	Neural	HMM
Event 1	02022	22222(60%)	00222(60%)	00220(40%)	02020(80%)
Event 2	22022	20020(60%)	02020(60%)	00022(60%)	02020(60%)
Event 3	00200	22020(20%)	10200(80%)	00220(80%)	00000(80%)
Event 4	20000	20000(100%)	20000(100%)	02102(20%)	20000(100%)
Event 5	20020	20000(80%)	02000(40%)	02002(20%)	20000(80%)
Event 6	22222	22020(60%)	00222(60%)	00022(40%)	22202(80%)

* The values in brackets are the prediction accuracy of the corresponding model, and the visual representation of the accuracy comparison of different algorithms is shown in Figure 8.

TABLE 8. Credibility test of HMM algorithm on the Weibo event attention state.

Events	Actual sequences	Calculated by HMM algorithm	Credibility
Event1	22222020020000200202020	22222020020000200120202020	96%
Event2	22220220000000220102022220	20220220000020220102020020	88%
Event3	22202000022000202002000020	22200001020000220002000020	80%
Event4	20000222020002022002200	20000222020022002000222201	80%
Event5	22200020022022000002000200	22000020012022000002202000	80%
Event6	20000020020220020000020020	20000020020122020202020022	80%

of HMM algorithm, we compare the results in this work with Grey system, ARIMA model and Neural network model. A Grey system refers to the system which lack information, that is, a limited amount of data to estimate the behavior of unknown systems [37]. The grey model is also time series prediction model based only on a set of the most recent data depending on the window size of the predictor. A neural network [40] is a complex network system formed by a large number of simple processing units (or neurons) widely interconnected. Generally, prediction includes three layers, input layer, hidden layer and output layer. When neural network is being trained or operating normally, patterns of information are fed into the network via the input units, which trigger the layers of hidden units, and these in turn arrive at the output units. ARIMA, short for "AutoRegressive Integrated Moving Average" is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values. An ARIMA model can be viewed as a "filter" that tries to separate the signal from the noise, and the signal is then extrapolated into the future to obtain forecasts [39].

For each case data in this work is part into 5:1 in chronological order, the first 5/6 part is used as the training set, and the other is the test set, to carry out the results of the Weibo event attention trend prediction, shown in Table 7. By the simulation results obtained above, the accuracy of the prediction results of the HMM prediction model, Grey prediction model, ARIMA prediction model and Neural network prediction model proposed in this article are compared and analyzed, as shown in Figure 8.

It can be seen from Figure 8 that prediction results of ARIMA model and HMM model are relatively stable, but the former is lower than that of HMM. Therefore, the prediction results of the HMM based on user popularity established in this paper are better than those of the three, and are closer to the real change state of Weibo events attention.

TABLE 9. Event type, time and data volume.

Events	Events	Data Time	Event type	|microblogs|
Event 1	Xi'an Benz Rights Protection	2019.4.12-2019.5.12	Social management	88149
Event 2	Tianlin Zhai academic fraud	2019.2.8-2019.3.8	Entertainment	6465
Event 3	African locust plague	2020.2.2-2020.3.2	Disaster	14791
Event 4	2020 college entrance examination postponed	2020.3.31-2020.4.30	Education	5353
Event 5	Fang Fang's Diary	2020.5.1-2020.5.31	Public health	28551
Event 6	Teacher salary is not lower than civil servant	2020.5.19-2020.6.19	Public management	13140

$*$ |microblogs| : the number of microblogs related to the Weibo event

TABLE 10. Statistical properties of data.

Subject	Classification	Properties	Event1	Event2	Event3	Event4	Event5	Event6
Microblog	original microblog	|original|	35102	5210	6808	5158	10108	10675
		N_{theme}	28975	8906	11111	14735	13362	26322
		$N_{mention}$	4433	435	439	114	2733	534
		N_{link}	4127	204	1220	279	657	777
	retweet microblog	|retweet|	53047	1255	7983	195	18443	2465
		N_{theme}	29937	1784	1134	202	12551	4472
		$N_{mention}$	6141	965	6011	23	25575	917
		N_{link}	4181	12	119	34	275	43
User	Authen	|Athn|	13612	1030	2558	1925	3060	4052
		|ave-flw|	1719169	472256	969227	1857106	365532	215864
	non-Athn	|non-Athn|	54813	4514	10713	2947	11436	6574
		|ave-flw|	47176	661	10377	34904	4249	5452

$*$ | $*$ |: the number of $*$.

$*$ $Athn$ and $ave-flw$ are authenticated and average-followers for short, respectively.

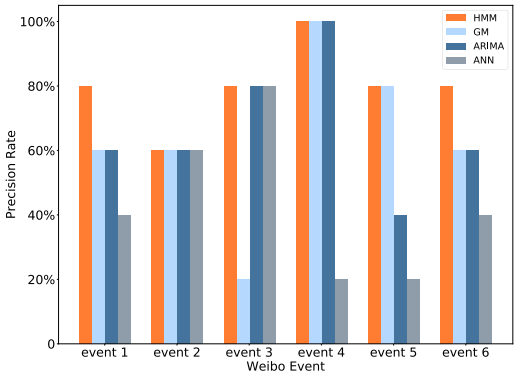


FIGURE 8. Comparative analysis of predicted results. HMM is a Hidden Markov Model, GM is a grey prediction model, ARIMA is an autoregressive moving average forecasting model, and ANN is artificial neural network prediction. The vertical axis represents the accuracy of the prediction results, and the horizontal axis represents six events. Four different color bars on each event represent the results of four different algorithms: the orange bar is the HMM algorithm in this paper, the light blue bar is GM, the dark blue bar is ARIMA, and the gray bar is ANN.

With the rapid development of Weibo, it has become the center of public opinion information dissemination in China. A large amount of information is generated every day, and mixed with a lot of bad information, which brings challenges to the Internet content governance. To ensure the implementation of Internet content governance, in this research, we propose a method that combines the influencing factors of the user and the Weibo event itself to predict the development trend of future events. First, we put forward some indicators to evaluate the popularity of users through the users themselves and the microblog information published by users. Then, by analyzing the correlation between user popularity and Weibo events attention, using the logical framework of Hidden Markov Model, a prediction algorithm is proposed to predict the future trend of events and achieve better accuracy.

Contributions in this work for Weibo events evaluation include: the first is the user popularity measurement defined by combining comprehensively the characteristics of microblogs and users; the second is the definition of Weibo event attention degree of original microblogs and retweet microblogs; the third is to predicate the trends of event attention by an algorithm based on Hidden Markov Model, where the observation sequence and the hidden state sequence are represented by user popularity and Weibo event attention, respectively.

Finally, we summarize the limitations of the work and look forward to the future work plan. (1) From the perspective of data type, there are many kinds of information carriers in Sina Weibo platform, including video, picture, text, audio

and other forms. However, in our research, we only consider the text information. If an event exists more in the form of other complex carriers, the measurement of event attention will be inaccurate. Therefore, in the following research, it is meaningful to carry out the quantitative research of event attention considering multiple types of information carriers, which can quantify the event attention from a more comprehensive perspective. (2) We try our best to avoid the collinearity between the calculated parameters when designing the evaluation index of user popularity. However, in the process of Weibo event development, it may be affected by the external environment and change dramatically. Even if we use HMM, which has the inherent advantages of short-term prediction, as the logical framework of prediction, when the event popularity changes suddenly due to external interference, the prediction accuracy of the model may be affected. Therefore, in the future research, how to design the prediction model to adapt to the external environment is worth further thinking. (3) In the quantitative index of Weibo event attention, we consider the influence of several actual parameters. However, the heterogeneity of events, such as the different nature of different events, different user groups talking about events and other factors, will also interfere with the prediction results. Although we used six different events to test the impact of different events on our algorithm, we did not study it further. In our follow-up work, we plan to analyze the commonness of different events from the perspective of users and microblog content, and design more reasonable evaluation index of the user popularity, so as to improve the accuracy of prediction.

APPENDIX 1

Six different types of events, including social management, entertainment, disaster, public health, education, and public management are shown in Table 9.

Table 10 shows the statistical characteristics of microblogs and users. For microblogs, such as the number of original microblog and retweet microblog, the number of topics, the number of @, and the number of out-links about different microblog types for different events. For users, including the number of authenticated users and non-authenticated users, and the average number of followers of different types of users.

Table 11 shows the specific data of the total number of microblog posted per hour of the 6 Weibo events. Table 12 and Table 13 show the state transition matrix A of other event attention and the output probability matrix B of user popularity.

TABLE 11. The total number of posts per hour of 6 Weibo events.

T_w	N	T_w	N	T_w	N
0:00-1:00	4655	8:00-9:00	6646	16:00-17:00	8314
1:00-2:00	2926	9:00-10:00	12145	17:00-18:00	7627
2:00-3:00	1962	10:00-11:00	11373	18:00-19:00	6406
3:00-4:00	1305	11:00-12:00	10630	19:00-20:00	5550
4:00-5:00	1408	12:00-13:00	9830	20:00-21:00	6750
5:00-6:00	1825	13:00-14:00	7072	21:00-22:00	7574
6:00-7:00	3646	14:00-15:00	6913	22:00-23:00	6268
7:00-8:00	4606	15:00-16:00	6569	23:00-24:00	5924

TABLE 12. Attention state transition probability matrix A of Weibo events.

Event	EA in S	up	unchange	down
Event 2	up	0.33210	0.08363	0.58427
	unchange	0.99933	0.00033	0.00034
	down	0.72718	0.00033	0.27249
Event 3	up	0.66687	0.06671	0.26642
	unchange	0.99933	0.00033	0.00034
	down	0.62473	0.00034	0.37493
Event 4	up	0.53829	0.07719	0.38452
	unchange	0.00033	0.00033	0.99934
	down	0.54535	0.00033	0.45432
Event 5	up	0.64282	0.07168	0.28550
	unchange	0.00033	0.00033	0.99934
	down	0.66653	0.00033	0.33134
Event 6	up	0.38406	0.07714	0.53880
	unchange	0.00033	0.00033	0.99934
	down	0.79986	0.00033	0.19981

TABLE 13. The probability matrix B of estimating visible state of Weibo events.

Event	Y_u in O	up of EA	unchange of EA	down of EA
Event 2	up of Y_u	0.99894	0.00033	0.00033
	unchange of Y_u	0.00276	0.99643	0.00081
	down of Y_u	0.00133	0.00056	0.99811
Event 3	up of Y_u	0.99856	0.00053	0.00091
	unchange of Y_u	0.00091	0.99786	0.00123
	down of Y_u	0.00086	0.00047	0.99867
Event 4	up of Y_u	0.99867	0.00042	0.00091
	unchange of Y_u	0.00157	0.99707	0.00136
	down of Y_u	0.00084	0.00041	0.99875
Event 5	up of Y_u	0.99867	0.00040	0.00093
	unchange of Y_u	0.00089	0.99752	0.00159
	down of Y_u	0.00079	0.00038	0.99883
Event 6	up of Y_u	0.99881	0.00042	0.00077
	unchange of Y_u	0.00077	0.99848	0.00075
	down of Y_u	0.00102	0.00038	0.99860

TABLE 14. The number of retweets of a single microblog by users in different categories.

User	Authentication	class	followers	Retweets
User 1	Grassroots users	<500	290	4
User 2	Enterprise users	500-5000	4052	14
User 3	Comic users	5000-50000	21151	156
User 4	News users	>50000	76663001	5126
User 5	Celebrity users	>50000	1913297	855
User 6	Games users	>50000	10991841	2835
User 7	Media users	>50000	1066382	4066
User 8	Marketing users	>50000	181405	43

An user who retweet microblog is called as a follower, and all followers are parted into 4 classes by their sizes, shown in Table 14.

The correlation of the Weibo event attention and the user popularity is modeled as a polynomial regression equation based on dynamic coefficients, shown in Table 15, where Multiple R, the R^2, and the Adjusted R^2 are used to judge the fitness of the polynomial regression model. Table 15 shows that the polynomial regression model has a good fitness.

TABLE 15. Fitting $Y_t = a_1 X_{t-1}^3 + a_2 X_{t-1}^2 + a_3 X_{t-1} + a_4, t > 1$ by Weibo events, where Y_t is the Weibo event attention at time t, and X_{t-1} is the user popularity at time $t-1$.

Events	a_1	a_2	a_3	a_4	Multiple R	R^2	Adjusted R^2
Event1	0.0011	-0.1541	17.121	-6.203	0.998	0.997	0.996
Event2	2.0562	-15.925	48.992	-8.061	0.943	0.890	0.886
Event3	-0.0335	1.0125	1.996	9.178	0.987	0.975	0.974
Event4	-0.0014	0.168	11.286	0.538	0.999	0.999	0.998
Event5	0.1043	-1.8702	16.750	-6.815	0.924	0.853	0.847
Event6	0.0048	0.3502	20.059	-3.072	0.989	0.979	0.978

TABLE 16. When 60% data is used to predict 40% data, the performance of different prediction algorithms is compared.

Events	Actual	Grey	ARIMA	Neural	HMM
Event 1	202022002022	000000000000(33.3%)	202012000222(75.0%)	020020200202(33.3%)	222000222222(50.0%)
Event 2	202222022022	200200000002(41.6%)	202020212000(25.0%)	200202000022(41.6%)	200200000002(41.6%)
Event 3	022000200020	200000000000(58.3%)	020012002000(50.0%)	022002022002(50.0%)	200000000000(58.3%)
Event 4	200220020000	200200200000(75.0%)	022220122002(50.0%)	022020200120(33.3%)	200000000000(75.0%)
Event 5	200020020020	000000100101(41.6%)	022020001002(41.6%)	000200222220(50.0%)	222200000000(50.0%)
Event 6	000202022222	202002000000(33.3%)	220200200201(33.3%)	002000202002(41.6%)	202020220202(33.4%)

* The values in brackets are the prediction accuracy of the corresponding model.

TABLE 17. When 70% data is used to predict 30% data, the performance of different prediction algorithms is compared.

Events	Actual	Grey	ARIMA	Neural	HMM
Event 1	022002022	202002000(55.5%)	212000002(55.5%)	022200220(55.5%)	200000020(44.5%)
Event 2	222022022	020110100(11.1%)	202012000(55.5%)	020202200(22.2%)	202020202(55.5%)
Event 3	002000200	200200000(55.5%)	212002000(55.5%)	020020200(66.7%)	200000000(66.7%)
Event 4	220020000	000000000(66.7%)	000000000(66.7%)	002202020(22.2%)	200000000(77.8%)
Event 5	020020020	200000000(55.5%)	200012220(44.5%)	002002022(44.5%)	220000000(66.7%)
Event 6	202022222	000000000(22.2%)	200200202(33.3%)	002200022(44.5%)	202020202(77.8%)

* The values in brackets are the prediction accuracy of the corresponding model.

APPENDIX 2

In the text, the original data is divided into six equal parts according to the amount of data and time order, and the first five parts as the training set are used to predict the last one. In this way, the user attention shown in 80% of the data is used to predict the event attention implied in 20% of the data. In order to further compare the impact of training set size on the prediction results, we supplement two groups of control experiments, which use 60% of the data to show the user attention to predict the event attention implied in the remaining 40% of the data, and use 70% of the data to predict the remaining 30% of the data. The test results are shown in Table 16 and Table 17 respectively. Combining with Table 7, it can be found that with the decrease of the amount of predicted data, that is, the decrease of the predicted time, the prediction performance of our method is improved more obviously than that of other algorithms, which means that our algorithm has better performance for short-term prediction. In addition, even with different training set sizes, our algorithm still performs well.

REFERENCES

[1] P. Yang, G. Yang, J. Liu, J. Qi, Y. Yang, X. Wang, and T. Wang, "DUAPM: An effective dynamic micro-blogging user activity prediction model towards Cyber-Physical-Social systems," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5317–5326, Aug. 2020.

[2] B. Wu and H. Shen, "Analyzing and predicting news popularity on Twitter," *Int. J. Inf. Manage.*, vol. 35, no. 6, pp. 702–711, Dec. 2015.

[3] Z. Wang and Y. Guo, "Empower rumor events detection from Chinese microblogs with multi-type individual information," *Knowl. Inf. Syst.*, vol. 62, no. 9, pp. 3585–3614, Sep. 2020.

[4] L. DeNardis and A. M. Hackl, "Internet governance by social media platforms," *Telecommun. Policy*, vol. 39, no. 9, pp. 761–770, Oct. 2015.

[5] S. Zhang and Q. Lv, "Hybrid EGU-based group event participation prediction in event-based social networks," *Knowl.-Based Syst.*, vol. 143, pp. 19–29, Mar. 2018.

[6] J. Zhao, W. Wu, X. Zhang, Y. Qiang, T. Liu, and L. Wu, "A short-term trend prediction model of topic over Sina Weibo dataset," *J. Combinat. Optim.*, vol. 28, no. 3, pp. 613–625, Oct. 2014.

[7] Q. Cao, H. Shen, H. Gao, J. Gao, and X. Cheng, "Predicting the popularity of online content with group-specific models," in *Proc. 26th Int. Conf. World Wide Web Companion (WWW Companion)*, 2017, pp. 765–766.

[8] T. Liu, Y. Zhong, and K. Chen, "Interdisciplinary study on popularity prediction of social classified hot online events in China," *Telematics Informat.*, vol. 34, no. 3, pp. 755–764, Jun. 2017.

[9] L. Hong, O. Dan, and B. D. Davison, "Predicting popular messages in Twitter," in *Proc. 20th Int. Conf. Companion World Wide Web (WWW)*, 2011, pp. 57–58.

[10] H. Cui and J. Kertész, "Attention dynamics on the Chinese social media Sina Weibo during the COVID-19 pandemic," *EPJ Data Sci.*, vol. 10, no. 1, pp. 1–16, Dec. 2021.

[11] H. Ma, W. Qian, F. Xia, X. He, J. Xu, and A. Zhou, "Towards modeling popularity of microblogs," *Frontiers Comput. Sci.*, vol. 7, no. 2, pp. 171–184, Apr. 2013.

[12] X. Lu, Z. Yu, B. Guo, and X. Zhou, "Predicting the content dissemination trends by repost behavior modeling in mobile social networks," *J. Netw. Comput. Appl.*, vol. 42, pp. 197–207, Jun. 2014.

[13] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," in *Proc. AAAI Conf. Weblogs Social Media*, vol. 14, 2010, pp. 10–17.

[14] M. Cataldi, N. Mittal, and M.-A. Aufaure, "Estimating domain-based user influence in social networks," in *Proc. 28th Annu. ACM Symp. Appl. Comput. (SAC)*, 2013, pp. 1957–1962.

[15] J. Wei, G. Mengdi, W. Xiaoxi, and W. Xianda, "A new evaluation algorithm for the influence of user in social network," *China Commun.*, vol. 13, no. 2, pp. 200–206, 2016.

[16] M. Garg and M. Kumar, "Review on event detection techniques in social multimedia," *Online Inf. Rev.*, vol. 40, no. 3, pp. 347–361, Jun. 2016.

[17] C. Francalanci and A. Hussain, "Influence-based Twitter browsing with NavigTweet," *Inf. Syst.*, vol. 64, pp. 119–131, Mar. 2017.

[18] M. Lu, Z. Wang, and D. Ye, "Topic influence analysis based on user intimacy and social circle difference," *IEEE Access*, vol. 7, pp. 101665–101680, 2019.

[19] R. Dong, L. Li, Q. Zhang, and G. Cai, "Information diffusion on social media during natural disasters," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 265–276, Mar. 2018.

[20] Z. Zengin Alp and Ş. G. Ögüdücü, "Identifying topical influencers on Twitter based on user behavior and network topology," *Knowl.-Based Syst.*, vol. 141, pp. 211–221, Feb. 2018.

[21] J. Y. M. Nip and K.-W. Fu, "Challenging official propaganda? Public opinion leaders on Sina Weibo," *China Quart.*, vol. 225, pp. 122–144, Mar. 2016.

[22] K. K. Chakravarthi, L. Shyamala, and V. Vaidehi, "Budget aware scheduling algorithm for workflow applications in IaaS clouds," *Cluster Comput.*, vol. 23, no. 4, pp. 3405–3419, Dec. 2020.

[23] Y. Ma, J. He, and Q. Yu, "Modeling on social popularity and achievement: A case study on table tennis," *Phys. A, Stat. Mech. Appl.*, vol. 524, pp. 235–245, Jun. 2019.

[24] G. Wu, D. Kaifeng, Z. Jian, Z. Xianbo, and T. Daizhong, "Integrated sustainability assessment of public rental housing community based on a hybrid method of AHP-entropy weight and cloud model," *Sustainability*, vol. 9, no. 4, pp. 603–627, 2017.

[25] M. Peng, J. Zhu, H. Wang, X. Li, and Y. Zhang, "Mining event-oriented topics in microblog stream with unsupervised multi-view hierarchical embedding," *ACM Trans. Knowl. Discovery Data*, vol. 12, no. 3, p. 38, 2018.

[26] X. Feng, Q. Zhao, J. Ma, and G. Jiang, "On modeling and predicting popularity dynamics via integrating generative model and rich features," *Knowl.-Based Syst.*, vol. 196, May 2020, Art. no. 105786.

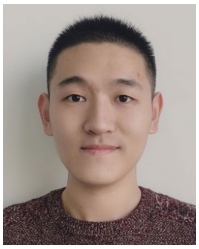
[27] B. Rezaie, M. Zahedi, and H. Mashayekhi, "Measuring time-sensitive user influence in Twitter," *Knowl. Inf. Syst.*, vol. 62, no. 9, pp. 3481–3508, Sep. 2020.

[28] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," in *Proc. Int. Conf. Web Search Web Data Mining (WSDM)*, 2008, pp. 207–218.

[29] F. Wu and B. A. Huberman, "Novelty and collective attention," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 45, pp. 17599–17601, Nov. 2007.
[30] Z. Wang, H. Liu, W. Liu, and S. Wang, "Understanding the power of opinion leaders' influence on the diffusion process of popular mobile games: Travel Frog on Sina Weibo," *Comput. Hum. Behav.*, vol. 109, Aug. 2020, Art. no. 106354.
[31] X. Chen, R. Geng, and S. Cai, "Predicting microblog users' lifetime activities–A user-based analysis," *Electron. Commerce Res. Appl.*, vol. 14, no. 3, pp. 150–168, 2015.
[32] K. R. Žalik and B. Žalik, "Validity index for clusters of different sizes and densities," *Pattern Recognit. Lett.*, vol. 32, no. 2, pp. 221–234, Jan. 2011.
[33] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, Nov. 1987.
[34] T. Chadza, K. G. Kyriakopoulos, and S. Lambotharan, "Analysis of hidden Markov model learning algorithms for the detection and prediction of multi-stage network attacks," *Future Gener. Comput. Syst.*, vol. 108, pp. 636–649, Jul. 2020.
[35] S. Liu, K. Zheng, L. Zhao, and P. Fan, "A driving intention prediction method based on hidden Markov model for autonomous driving," *Comput. Commun.*, vol. 157, pp. 143–149, May 2020.
[36] L. Wang, Q. Chen, Y. Sun, Y. Ying, X. Shi, and L. Fu, "An active fault-tolerant control method based on moving window hidden Markov model," *Chem. Eng. Sci.*, vol. 227, Dec. 2020, Art. no. 115865.
[37] J. L. Deng, "Introduction to grey system theory," *J. Grey Syst.*, vol. 1, no. 1, pp. 1–24, 1989.
[38] L. H. Holley and M. Karplus, "Protein secondary structure prediction with a neural network," *Proc. Nat. Acad. Sci. USA*, vol. 86, no. 1, pp. 152–156, 1989.
[39] J. Penzer and B. Shea, "Finite sample prediction and interpolation for ARIMA models with missing data," *J. Forecasting*, vol. 18, no. 6, pp. 411–419, Nov. 1999.
[40] M. A. Ghorbani, R. C. Deo, S. Kim, M. Hasanpour Kashani, V. Karimi, and M. Izadkhah, "Development and evaluation of the cascade correlation neural network and the random forest models for river stage and river flow prediction in Australia," *Soft Comput.*, vol. 24, no. 16, pp. 12079–12090, Aug. 2020.

YINGHONG MA was born in Shandong, China. She received the B.S. degree in mathematics from Shandong Normal University, in 1996, and the M.S. and Ph.D. degrees in mathematics from Shandong University, China, in 1999 and 2002, respectively. In recent years, she has been engaged in the research of management decision theory and methods, and complex network theory and application in scientific research.

HUI JIAO received the B.S. degree from Shandong Normal University, in 2018, where she is currently pursuing the Graduate degree. Her research interests include information dissemination and public opinion management.



LE SONG received the B.S. and M.S. degrees from Shandong Normal University, in 2015 and 2018, respectively, where he is currently pursuing the Ph.D. degree. His research interests include information dissemination, scientific evaluation, and intelligent computing.