

Research Article

Information Propagation Prediction Based on Key Users Authentication in Microblogging

Miao Yu , Yongzheng Zhang , Tianning Zang, Yipeng Wang, and Yijing Wang

Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

Correspondence should be addressed to Yongzheng Zhang; zhangyongzheng@iie.ac.cn

Received 1 September 2017; Revised 6 December 2017; Accepted 25 December 2017; Published 6 February 2018

Academic Editor: Qi Jiang

Copyright © 2018 Miao Yu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In microblogging, key users are a significant factor for information propagation. Key users can affect information propagation size while retweeting the information. In this paper, to predict information propagation, we propose a novel linear model based on key users authentication. This model mines key users to dynamically improve the linear model while predicting information propagation. So our model can not only predict information propagation but also mine key users. Experimental results show that our model can achieve remarkable efficiency on predicting information propagation problem in real microblogging networks. At the same time, our model can find the key users who affect information propagation.

1. Introduction

In the past, when we mentioned authentication, we thought of password research first. A mass of research [1–5] focuses on passwords mechanism and other authentication mechanisms for user authentication in various computer systems. But recently, with the rapid development of microblogging, the scale of users is becoming larger and larger. It plays an important role as an efficient media for fast spreading information, ideas, and influence among huge population. In microblogging users play different roles in information propagation. The immense popularity of microblogging provides great opportunities for social network public opinion [6, 7]. At the same time, in microblogging, participants have the characteristic of high dynamics, self-organization, and heterogeneity. All of the above factors make the dissemination and evolution process of network opinions become more random and complicated. Due to the above challenges, we should focus on microblogging user authentication of role in information propagation.

Information propagation prediction can be identified as an early warning scheme for controlling large-scale outbreak of negative network public opinion in microblogging. In previous researches on information propagation prediction, information propagation prediction can be divided

into microprediction and macroprediction. Microprediction [8–11] mainly does research on user's retweeting a message. Macroprediction mainly focuses on predicting the macroindicator of information, such as scale, depth, and speed. In this paper, we mainly do research on macroprediction about the scale prediction of information propagation. In the previous researches of macropredictions, users are not adequately considered as a key factor for the scale prediction information propagation. Existing researches [11–13] mainly consider information attribute and user attribute in predicting information propagation. But extensive research [14–16] shows that the large propagation of social network public opinions is usually caused by one user or multiple users. These users are called key users. The key user can affect the prediction result when prediction time windows show key user.

Due to the above challenges, we propose a novel linear model based on key users authentication to predict information propagation. This model considers the influence of key user for information propagation prediction. We mine key users in process of information prediction. Then we add key user function into linear model. Hence, this model can not only predict information but also mine key user. To evaluate the performance of our model, we conduct extensive experiments on microblogging datasets. Experimental results

demonstrate that our model is high in accuracy in comparison with the baseline algorithms. At the same time, our model can find key users in the process of information propagation prediction.

The remainder of this paper is organized as follows: Section 2 introduces some features of information propagation in microblogging. Section 3 proposes a novel linear model to predict information propagation and mine key user in microblogging. The proposed model is validated through experiments in comparison with other baseline algorithms in Section 4. Section 5 gives a conclusion and directions for the future works.

2. Analysis of Information Propagation in Microblogging

In this section, the important characteristics of information propagation are analyzed for predicting information propagation in microblogging.

2.1. Analysis of Time Factor. In microblogging time cyclical factor is a key factor for information propagation. According to the Sina Weibo Data Center report published in 2015 shows that microblogging users have high dependence for Sina Weibo. The number of messages posted by microblogging users starts to rise from 5 a.m. and reaches the first peak at 12, followed by a little downgrade between 1 and 2 p.m., but this number rises mildly after 3, after that from 7 it rises smoothly again, while it begins to go down until it reaches a second peak at about 10 in the evening. Thus, Sina Weibo plays an important role in the daily network life.

In this paper, we find that retweeting and commenting behavior also satisfy this rule by analyzing Sina Weibo data. User behavior basically accords with people's daily life. As shown in Figure 1, Sina Weibo shows that user number reaches a peak at about 10 p.m. per day, and the trough period at about 4 a.m. per day. Hence, information diffusion can be affected by user's daily life.

In fact, the information propagation of opinion public can last for a long time. So user's daily life must be considered in predicting information diffusion. To reduce the effect of daily life, we use time cycle weight to process microblogging data. The time cycle weights are described in Table 1.

2.2. Analysis of Information Propagation Structure Factor. In the propagation of microblogging, the propagation structure of information is an important metric of estimating information influence. The higher information influence, the more complex propagation structure. In literature [17], the author carried out the statistics about the height of information propagation tree, where the common propagation height is one, about 95.8%, and the longest propagation link is eleven hops. The propagation structure is usually complex in the real hot information propagation, which is not fit for the information propagation. Hence, in order to reduce the difference from information structure on predictions, we propose a novel approach where information propagation

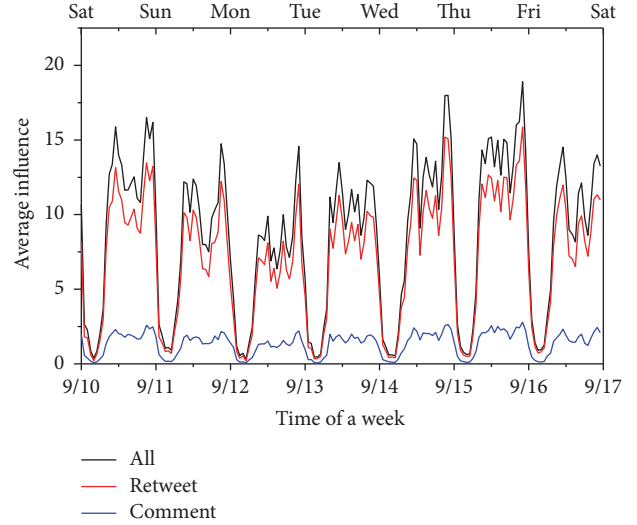


FIGURE 1: Illustration of time cycle.

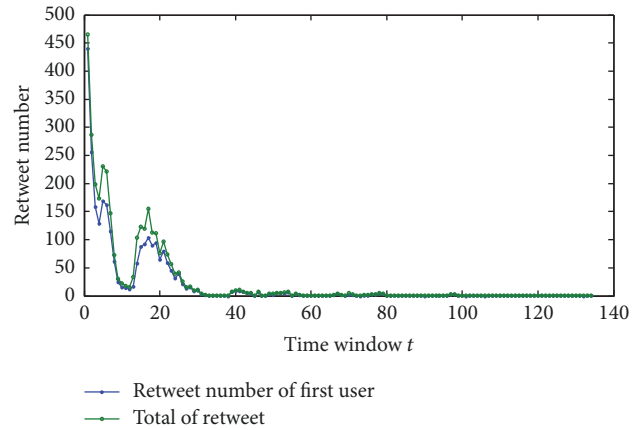


FIGURE 2: Diffusion scale of microblogging message.

scale is divided into multiple one-layer information propagation trees. While the first-layer retweet sale is helpful to predict the information propagation. The further propagation effectively shows the characteristics of information propagation. We take a detailed analysis on a microblogging message from Sina Guangzhou about Shenzhen landslides, and the microblogging message can be divided into multiple parts with one-hour time window, as shown in Figure 2.

From the above data analysis, we can observe that, in the process of information propagation, the scale of messages propagation is similar to the scale of first-layer propagation of messages, the scale of direct retweet of original messages. In the following prediction approach, the publisher of messages is added into model as the first key. At the same time, the whole propagation of the message is shown in Figure 3 that is made by microblogging visual analysis tools, PKUVIS.

Figure 3 shows the propagation structure of one message. From the figure, we can observe that the propagation structure of one message in microblogging is very complex. The information has been propagated to many levels. If the

TABLE 1: Time cycle weight.

0:00~0:59	1:00~1:59	2:00~2:59	3:00~3:59	4:00~4:59	5:00~5:59
3	2	1	1	1	1
6:00~6:59	7:00~7:59	8:00~8:59	9:00~9:59	10:00~10:59	11:00~11:59
2	3	4	4	4	4
12:00~12:59	13:00~13:59	14:00~14:59	15:00~15:59	16:00~16:59	17:00~17:59
4	4	4	4	4	4
18:00~18:59	19:00~19:59	20:00~20:59	21:00~21:59	22:00~22:59	23:00~23:59
4	5	5	5	5	4



FIGURE 3: Diffusion structure of microblogging message.

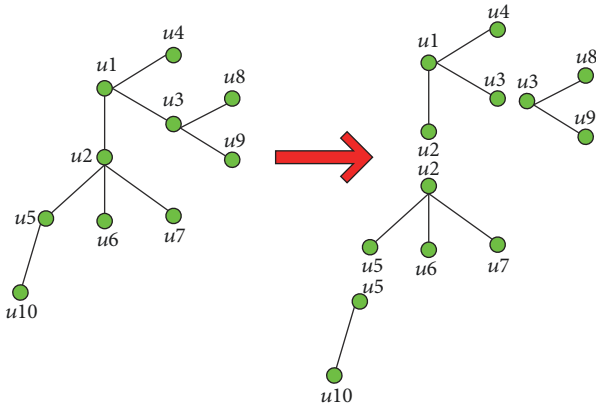


FIGURE 4: Splitting information diffusion tree.

propagation structure is considered on the propagation prediction, the prediction of messages would be more complex. In order to avoid the influence of propagation structure on the propagation prediction, we consider each user as an independent unit to count the number of retweets. So we divide the propagation tree of one message into multiple one-layer propagation tree sets. The detailed process is shown in Figure 4.

As shown in Figure 4, we only consider the direct retweet number of each user. A complex propagation tree is divided into multiple one-layer propagation trees. In one-layer propagation tree, the leaf node represents the direct retweet number of the root user. Thus the retweet number

of one message N_t^m can be represented as the following equation:

$$N_t^m = \sum_{u \in UN_t^m} N_t^{(m,u)}, \quad (1)$$

where N_t^m is the number of retweeted messages m at time t ; $N_t^{(m,u)}$ is the number of retweeted messages m before time t ; UN_t^m is the users set of messages m before time t .

Therefore, the message propagation prediction can be transformed into a novel problem that can predict multiple-user retweet number with m message.

2.3. Analysis of Key User Factor. In the research on social networks, key users mining is always an important, hot research issue. The researchers carry on back analysis on the microblogging information dissemination through analyzing key users. At the same time, the key user is also an important factor that cause information propagation. Large-scale of information propagation, even secondary burst, is usually caused by some key users. The key users have great effect on message propagation. Hence, it is necessary to consider the effect of users on the prediction of messages. In the section, we will detail the effect of key users on the propagation of messages.

In the previous section, a message publisher as message creator can be considered as the first key user. When other key users do not appear, the propagation scale of messages is only related to message publishers. However, when other key users appear, the retweet scale of messages is affected by key users. We given an example that the official microblogging of ‘‘CCTV news’’ published ‘‘senior girls by Harvard College enrolment in advance.’’ We compare the direct retweet of publishers with the retweet number of whole messages, as shown in Figure 5.

As shown in Figure 5, ‘‘CCTV News’’ publishes the message, as the creator and the first key user of the message. Before time $t = 48$, the propagation scale of messages is almost the same with the direct retweet scale of messages creator. At time $t = 48$, the user ‘‘shibugui’’ retweets the message, which causes the second emerging retweet number. The user ‘‘Shi Bugui’’ can be considered as the second key user. When it is needed to predict the retweet number at time $t = 48$, the retweet scale can not be predicted because the users can not be determined at time $t = 48$. At time $t = 48$, the retweet of key user ‘‘Shi Bugui’’ causes the inaccuracy

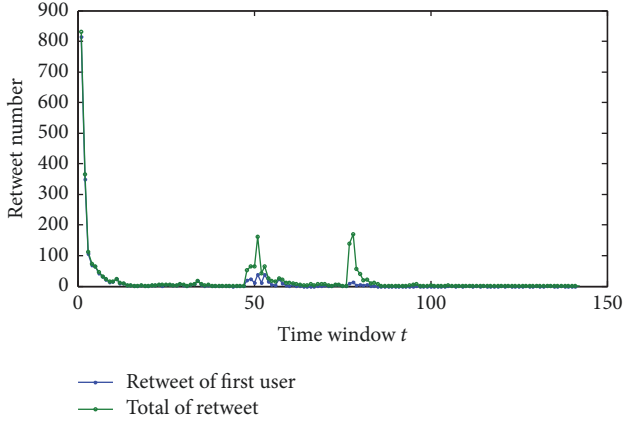


FIGURE 5: Appearance time of key users in information diffusion.

of prediction model. At time $t = 50$, the third and fourth key users “Wang Xiaoshan” and “Kou Ergou” retweet the message, which cause the message to be further retweeted. After the time $t = 50$, the messages fade more and more. At time $t = 77$, the fifth key user of message propagation “Gu Community” causes the fourth large retweet. So, from message sequence diagram, we can observe when key users take part in message propagation, key users are also a key factor that affects the prediction of messages propagation, which will be detailed in the following section.

3. Linear Prediction Model

In this section, we propose an efficient dynamic linear model to address the problem of information propagation prediction.

3.1. Influence of Key Users. A linear prediction model based on key users is introduced in the above context. In social networks, the information propagation is inextricably linked with users. How to formalize the information propagation of key users is key to the model. The information is posted or retweeted in microblogging, where the information propagation is decided by two factors. One is the influence of information itself that fades with the time. The other is the influence of users themselves that can affect the scale of propagation and speed of fading. According to the above reasons, literatures [18, 19] made researches on the information propagation model of users and put forward lots of equations that depict the influence of information propagation of users. Therefore, we define the influence of information propagation of users $R_u(t)$ as follows:

$$R_u(t) = c_u t^{-\alpha_u} e^{-t/\beta_u}, \quad (2)$$

where c_u, α_u, β_u are waiting parameters.

In the initial stage of information propagation, that is, when $t \ll \beta_u$, $e^{-t/\beta_u} \approx 1$. The power rate decay function plays a domain part, and retweeted number is slightly affected by exponential decay function. When $t \gg \beta_u$, because the fading speed of exponential decay function is larger than that of

power rate decay function, the retweeted number is decided by exponential decay function, which improves the speed of message decay. For the same message in the information propagation, c_u can be considered as the influence of key users. The larger the influence, the more the amount of initial forwarding. α_u denotes the decay speed of user influence. The larger α_u , the faster decay speed of key users. β_u represents the life cycle of key users’ influence. The larger β_u , the longer lasting time of retweeted β_u . Further, $R_u(t) = c_u t^{-\alpha_u}$ can be employed to sample the fitting of key users’ propagation influence. In order to guarantee the waiting valuation parameters, the following function can be employed to minimize the waiting valuation parameters.

$$\begin{aligned} & \text{minimize} \quad \sum N_t^{l(m,u)} - R_{u_i}^m(t), \\ & \text{s.t.} \quad R_{u_i}^m(t) \geq 0. \end{aligned} \quad (3)$$

3.2. Linear Model Based on Key Users. Based on the analysis of information propagation, this section mainly focuses on how to employ a function to predict the whole process of information propagation. According to the previous section, the scale of retweeted messages can be represented by (1). Hence, in order to reduce the influence of structure of message propagation on the prediction of message propagation, it is needed to fit the retweeted messages and the retweeted scale for predicting the whole scale. According to the described method, we define $P_m(t)$ as the retweet scale of messages m as follows:

$$P_m(t) = \sum_{u_i \in UR_m^m} R_{u_i}^m(t), \quad (4)$$

where $R_{u_i}^m(t)$ is the predicting valuing of retweet number of a user at time t .

From the equation, we can observe that it needs to fit the function of each user who retweets message m for predicting the retweet scale of message m at time t , which is with high cost. The above section proves that a larger number of retweeted messages are not retweeted again, while the retweet scale of messages only is related to some key users. Hence, it is not necessary to fit user function that retweet messages in the propagation of messages. Aiming at the above problem, we put forward an improved prediction model which makes the prediction based on a linear model of key users, as shown in Figure 6.

As shown in Figure 6, the retweet scale prediction equation is defined as (5). Equation (5) includes three parts. Considering that influence of the messages creator as first key user is different from that of other users, we employ parameters a_t, b_t to distinguish two types of key users; then d_t can be used to adjust the influence of partial messages from other nodes.

$$P_m(t) = d_t + a_t \cdot R_{u_1}^m(t) + b_t \sum_{u_i \in K_t^m} R_{u_i}^m(t), \quad (5)$$

where d_t, a_t, b_t are parameters to be estimated and $R_{u_i}^m(t)$ is the predicting value of retweet number of retweet message of

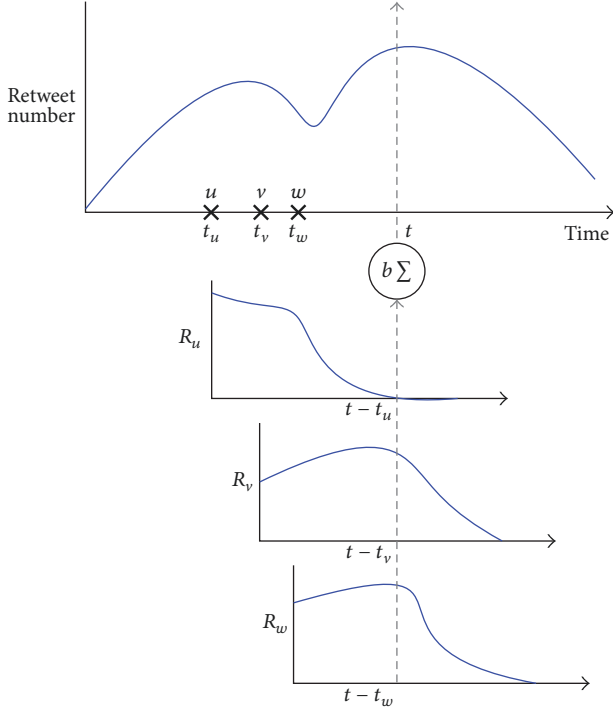


FIGURE 6: Linear model based on key users models the volume of microblogging message diffusion.

user u_i at time t . UR_t^m is the user set who retweet m message before time t .

Finally, the valuation parameters can be solved according to the formula with (3).

3.3. Mining Key User to Improve Model. In the propagation of information, the key users in training set are usually determined. However, when a key user appears in the prediction time window, the accuracy of prediction algorithm would be interfered, which reduces the accuracy of prediction. Hence, the section considers key users as an important factor for improving the accuracy of prediction for message retweet scale. When the sliding window includes a key user, the algorithm generates the corresponding prediction deviation. So, the section defines $K_m(t)$ as the threshold of key users existence:

$$K_m(t) = \frac{N_t^m - P_m(t)}{N_t^m}, \quad (6)$$

where N_t^m is the actual retweet value of message m at time t ; $P_m(t)$ is the predictive retweet number of messages m before time t .

If $K_m(t) < \theta$, the key user is not in the time window. If $K_m(t) \geq \theta$ and $N_t^m - P_m(t) \geq 10$, it is proved that the predicted time window may include key users that can interfere prediction. Therefore, it is needed to mine key users. $K_m(t) \geq \theta$ represents that there is large difference between prediction algorithm and real values, and $N_t^m - P_m(t) \geq 10$ for avoiding that the retweet scale with less than ten affects the computation. If $K_m(t) \leq -\theta$, it shows that some key users

in the previous window may be disabled, so it is needed to delete interference of some key users.

Firstly, the users in time window are sorted according to their retweet number. Then the users in the sorted set are successively added in the following equation until the following ruling condition is true.

$$\frac{N_t^m - P_m(t) + R_{u_i}^m(t)}{N_t^m} \leq \theta. \quad (7)$$

4. Experiments

In this section, we evaluate our method on real-world networks by comparing with current prediction models.

4.1. Experimental Setup

4.1.1. Baseline Algorithms. The compared algorithms include our KULM and two algorithms as follows:

- (1) MAM (moving average method): according to the time sequence, prediction window value equals the average of n windows before prediction windows.
- (2) ES (Exponential Smoothing): formula $S_t = \alpha y_t + (1 - \alpha)S_{t-1}$ for information diffusion prediction.

4.1.2. Datasets. Since there are no public datasets for information propagation prediction, we use data fetching program PKUVIS to obtain data on microblogging. The dataset includes seven messages of 6 hot topics, such as “landslide in shenzhen,” “wang baoqiang scold drunk driving,” and “cancel Late marriage leave.” The datasets of our experiment are described in Table 2.

All experiments are conducted on a hardware platform with CPU i5, 6 G memory, and 64-bit Win 7 system.

4.2. Experimental Analysis and Results. We evaluate our KULM on real-world Sina Weibo dataset by comparing with three baseline algorithms. In this paper, error rate is used as evaluation metrics. The error rate is computed as follows:

$$\text{error} = \frac{1}{n} \sum_{i=1}^n \frac{|P_m(t) - N_t^m|}{N_t^m}. \quad (8)$$

To evaluate the performance of KULM, we evaluate our method on Sina Weibo dataset by comparing with three baseline algorithms. A time window is 30 minutes, we use first ten windows as test data to predict follow-up of ten windows. The results of our experiment are described in Table 3.

As shown in Table 3, for the first key user, for ES model the lowest error rate is 26.0% and the highest is 59%. The model depends on datasets. Then MAM always has a high error rate, and the lowest error rate is only 40.4%. At last, KULM has the lowest error rate for predicting message propagation users on 7 types of microblogging messages. To prove that KULM can be used for mining key users at message propagation prediction, we demonstrate the effect of predicting and mining for No. 1 message. The result is shown in Figure 7.

TABLE 2: Dataset description.

Message number	Message user	Message topic
No. 1	Sina Guangdong	Landslides in Shenzhen
No. 2	Headline News	Landslides in Shenzhen
No. 3	Sina Entertainment	Baoqiang Wang inveighed against drunk driving escape
No. 4	Headline News	Late marriage leave cancel
No. 5	CCTV News	Senior three girls are admitted to Harvard
No. 6	CCTV News	Tramplng accident in Shanghai
No. 7	People's Daily	Landslides in Lishui, Zhejiang

TABLE 3: The compare predictive effect of algorithms.

Message number	Error rate			The existence of key user
	KULM	ES	MAM	KULM
No. 1	23.8%	42.1%	55.5%	No
No. 2	38.7%	59.4%	53.4%	No
No. 3	33.2%	46.2%	43.1%	No
No. 4	35.4%	39.2%	56.6%	Yes
No. 5	19.8%	35.6%	40.4%	No
No. 6	23.3%	26.0%	45.7%	Yes
No. 7	37.9%	45.2%	58.8%	Yes

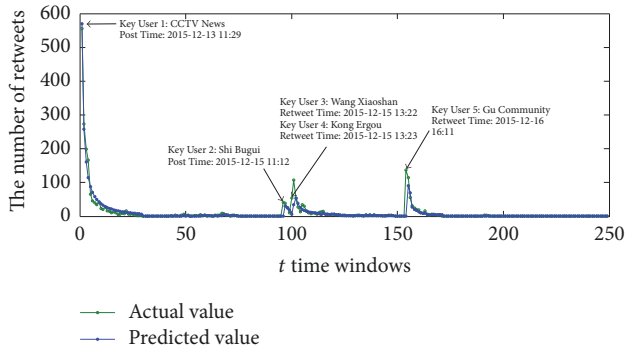


FIGURE 7: The result of predicting message propagation and mining key user by our model.

As shown in Figure 7, “CCTV News” publishes the message, as the creator and the first key user of message. Before time $t = 96$, the propagation scale of messages is almost similar with predicting propagation scale. When $t = 96$, there is much difference between the actual value and predicted value, so this time window may have key users. Then we mine key user in this window; we find the key user “Shi Bugui” who retweets the message at 11:12. At time windows $t = 100$, the time window shows much difference between the actual value and predicted value again. So we mine key user; then we find key users “Wang Xiaoshan” and “Kong Ergou.” The last key user is found at time window 154; the key user “Gu Community” retweets the message at 16:11 2015-12-16. Since the key user retweets message, actual value and predicted value make much different. So as shown in Figure 7, we can find clearly the arisen time of key users through difference value between actual value and predicted value. Hence our

KULM model can not only predict information propagation but also mine key users simultaneously.

5. Conclusion

In this paper, we consider the effect of key users on message propagation. Furthermore, we propose a novel linear model based on key users authentication to predict a message propagation in microblogging. This model can be improved dynamically through mining key users.

The experiments on real-world microblogging networks demonstrate the efficiency of our proposed algorithm for information propagation prediction in large-scale microblogging. The experiments also show that our algorithm is better than the baseline heuristic algorithms. And our algorithm can mine key users in the process of information propagation prediction.

In further works, we will consider more factors to improve KULM, such as the influence of topics. We will consider topic effect on information propagation by other topic messages. Our KULM will be used effectively to predict information under a specific topic.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by The National Key Research and Development Program of China under Grant no. 2016YFB0801304.

References

- [1] D. Wang, Z. Zhang, P. Wang, J. Yan, and X. Huang, "Targeted Online Password Guessing," in *Proceedings of ACM CCS 16*, pp. 1242–1254, Vienna, Austria, October 2016.
- [2] D. Wang, D. He, H. Cheng, and P. Wang, "FuzzyPSM: A new password strength meter using fuzzy probabilistic context-free grammars," in *Proceedings of the 46th IEEE/IFIP International Conference on Dependable Systems and Networks (DSN '16)*, pp. 595–606, France, July 2016.
- [3] C. Wang and G. Xu, "Cryptanalysis of three password-based remote user authentication schemes with non-tamper-resistant smart card," *Security and Communication Networks*, vol. 3, pp. 1–14, 2017.
- [4] R. Amin, S. K. H. Islam, M. K. Khan, A. Karati, D. Giri, and S. Kumari, "A two-factor RSA-based robust authentication system for multiserver environments," *Security and Communication Networks*, vol. 2017, Article ID 5989151, 15 pages, 2017.
- [5] Y.-H. Li and P.-J. Huang, "An Accurate and Efficient User Authentication Mechanism on Smart Glasses Based on Iris Recognition," *Mobile Information Systems*, vol. 2017, Article ID 1281020, 14 pages, 2017.
- [6] J. Wang, Z. Liu, and H. Zhao, "Micro-blogs entity recognition based on DSTCRF," *Journal of Electronics*, vol. 23, no. 1, pp. 147–150, 2014.
- [7] Z. Yang, K. Fan, Y. Lai, K. Gao, and Y. Wang, "Short texts classification through reference document expansion," *Journal of Electronics*, vol. 23, no. 2, pp. 315–321, 2014.
- [8] Z. Yang, J. Guo, K. Cai et al., "Understanding retweeting behaviors in social networks," in *Proceedings of the ACM Conference on Information and Knowledge Management*, pp. 1633–1636, October 2010.
- [9] H.-K. Peng, J. Zhu, D. Piao, R. Yan, and Y. Zhang, "Retweet modeling using conditional random fields," in *Proceedings of the 11th IEEE International Conference on Data Mining Workshops (ICDMW '11)*, pp. 336–343, Canada, December 2011.
- [10] T. R. Zaman, R. Herbrich, J. Van Gael, and D. Stern, "Predicting information spreading in twitter," in *Proceedings of the Workshop on computational social science and the wisdom of crowds*, pp. 17599–17601, 2010.
- [11] M. Wang, W. Zuo, and Y. Wang, "A multidimensional non-negative matrix factorization model for retweeting behavior prediction," *Mathematical Problems in Engineering*, vol. 2015, Article ID 936397, 10 pages, 2015.
- [12] A. Kupavskii, L. Ostroumova, A. Umnov et al., "Prediction of retweet cascade size over time," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, pp. 2335–2338, USA, November 2012.
- [13] J. Cheng, L. A. Adamic, P. A. Dow, J. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*, pp. 925–935, Republic of Korea, April 2014.
- [14] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto, "Finding trendsetters in information networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*, pp. 1014–1022, China, August 2012.
- [15] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: the million follower fallacy," in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM '10)*, pp. 10–17, May 2010.
- [16] X. Wu and J. Wang, "Micro-blog in China: Identify influential users and automatically classify posts on Sina micro-blog," *Journal of Ambient Intelligence and Humanized Computing*, vol. 5, no. 1, pp. 51–63, 2014.
- [17] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proceedings of the 19th International World Wide Web Conference (WWW '10)*, pp. 591–600, April 2010.
- [18] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM '10)*, pp. 599–608, IEEE, Sydney, Australia, December 2010.
- [19] C. X. Wang, X. H. Guan, T. Qin, and Y. . Zhou, "Modeling on opinion leader's influence in microblog message propagation and its application," *Journal of Software*, vol. 26, no. 6, pp. 1473–1485, 2015.

