Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 15, no. 11, November 2019, pp. 3024-3030
DOI: 10.23940/ijpe.19.11.p22.30243030

Mining Key Users of Microblog Topics based on Trust Model

Guozhong Dong[a], Bei Li[a], Xinhong Wei[a], and Tao Qin[b,*]

[a]*Department of Computer and Data Science, Henan University of Urban Construction, Pingdingshan, 467036, China*
[b]*National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijiing, 100190, China*

Abstract

Microblog topics play an important role in public opinion. The effective identification of key users in microblog topics is key for microblog public opinion mining. In this paper, we select topic-related microblog messages based on topic features. Candidate key users are selected according to the microblog topic user graph and edge weight. Key users of microblog topics are ranked according to the users' trust values. Experiments on Sina microblog topic datasets show that our method can mine high quality key users that lead to the formation of microblog topics in the early stage.

Keywords: trust model; microblog topics; key users mining

(Submitted on September 14, 2019; Revised on October 13, 2019; Accepted on November 25, 2019)

1. Introduction

Information dissemination has entered a new era of self-media with the rapid development of mobile Internet technology and the continuous popularization of intelligent terminal systems. The microblog is representative of the media in the Web 2.0 era. Its popularization, personalization, timeliness, and interactivity make it an important platform to obtain information, disseminate news, express opinions, and create public opinion. The emergence of microblog topics has accelerated the interaction between members of various groups, boosting personal emotions into social emotions. Sometimes, it directly promotes the progress of social security incidents.

The "2012 Internet Public Opinion Report" indicated that the microblog platform is the largest source of information for public emergencies, including "Beijing rainstorm disasters", "grab salt storms", "snoring Japanese cars", and other microblog topics. Microblog topics have attracted more and more attention from academic circles and the industry, and they are representative of the network's public opinions.

The low entry barrier of microblogs, the lack of responsibility, and the anonymous logins allow microblogs to accelerate the spread of negative energy sensation. The current system audit or manual real-time monitoring does not limit the spread of fake information due to the large amount of fragmented text information in microblogs. In addition to being a platform for the public to express their concerns and appeals, microblogs have also become a platform for the proliferation of fake information and rumors. Key users in microblog topics are the main reason for the formation of topics in microblogs. Therefore, effectively identifying high-quality key users in microblog topics is a research hotspot of microblog public opinion monitoring.

The main research objects of high-quality key user mining include microblog topic-related news and users. From the perspective of research objects, the challenges of high-quality key user mining problems in microblog topics include at least the following two points:

- The immediacy of the speed of message propagation: The speed of message dissemination is closely related to the type of users involved in message dissemination. The user type and user influence in the process of topic message

* Corresponding author.
E-mail address: qintao@cert.org.cn

diffusion are important indicators to measure the speed and scale of the topic. The rapid spread of news has created difficulties for mining key users who detonate microblog topics.

- The diversity of user quality: There is a large number of low-quality users in microblogs. These users promote the dissemination of topics by brushing and forwarding. This type of low-quality user participation in microblog topics poses a huge challenge for high-quality key users to mine microblog topics.

To address these challenges, this paper provides the following solutions:

- A user graph model is proposed for the characteristics of the rapid spread of message dissemination and the early formation of microblog topics. The model can select candidate key user sets and mine key users who promote the formation of microblog topics.
- A trust model is proposed for the diversity characteristics of microblog users' quality. By evaluating the interaction quality of candidate key users, the topic high-quality key users are explored.

2. Related Work

User influence has been extensively studied in the last decade [1-10]. Here, we introduce the studies most related to our work.

Brown et al. [1] investigated a modified k-shell decomposition algorithm based on user relationships to compute users' influence on Twitter. Cha et al. [2] analyzed the influence of Twitter users by employing three measures that capture different perspectives: indegree, retweets, and mentions. They found that influence is not determined by a single factor, but rather many factors. Fang et al. [3] developed a novel topic-sensitive influencer mining (TSIM) framework in interest-based social media networks to find topical influential users and images. Lee et al. [4] proposed a method to find influentials by considering both the link structure and the temporal order of information adoption in Twitter. Saez-Trumper et al. [5] put forward a ranking algorithm to detect trendsetters in information networks. The algorithm can identify people that spark the process of disseminating ideas that become popular in the network. Weng et al. [6] proposed an extension of the PageRank algorithm to measure the influence of users in Twitter, considering both the topical similarity between users and the link structure. Ye et al. [8] proposed a method combined with sentiment analysis to find suspicious as well as dominant users.

Compared with the above works, this paper introduces the time attribute of topic user behavior and trust model into the topic key user mining model, which can mine high-quality key users who promote the formation of microblog topics at an early stage.

3. Key User Mining Method based on Trust Model and User Graph Model

3.1. Candidate Key User Selection based on User Graph Model

The user graph of burst topic k can be formally defined as $G_k = < V_k, E_k, T_k >$. In detail, $V_k = \{u, \cdots v, \cdots\}$ is the user set of burst topic k. E_k represents the edge set among users, in which a directed edge (u,v) means that u is the follower of v. $T_k = \{t(u), \cdots, t(v), \cdots\}$ is the earliest post time set of users over burst topic k. By considering time information, the directed edges in the topic user graph model can represent the direction of information flow and play a key role in detecting key users. For each $(u,v) \in E_k$, the edge weight $w(u,v)$ and the normalization of edge weight $W(u,v)$ can be defined as follows:

$$w(u,v) = \begin{cases} e^{-\frac{t(u)-t(v)}{\delta}}, & \text{if } t(v) > 0 \text{ and } t(v) < t(u), \delta > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

$$W(u,v) = \frac{w(u,v)}{\sum_{p \in OUT_{G_k}(u)} w(u,p)} \tag{2}$$

Where $OUT_{G_k}(u)$ is the following set of user u and δ is the time adjustment factor. The user weight of user v in G_k, denoted by $UW(v)$, is given as follows:

$$UW(v) = dD(v) + (1-d) \sum_{p \in IN_{Gk}(v)} UW(p)W(p,v), \quad 0 \le d \le 1 \text{ and } D(v) = 1/|V_k| \tag{3}$$

Where d is the damping factor, $IN_{Gk}(v)$ is the follower set of user v, and $D(v)$ is a probability distribution over V_k. The distribution is topic dependent and is set to $1/|V_k|$ for all $v \in V_k$.

The candidate key users of burst topic k is formalized as CKU_k, which is defined as the top N user weight of users in V_k. We set N equal to 10 in this paper.

3.2. Key User Mining Method based on Trust Model

The user trust model of this paper considers two aspects: the global trust of microblog users and the trust between microblog users.

The global trust of microblog users is measured by the influence of microblog users in the microblog network. The global trust calculation formula is as follows:

$$GT_i = \frac{\sum_{m \in M(t)} R_m}{N_t} \tag{4}$$

Where GT_i is the global trust of microblog users i, R_m is the sum of the number of comments, the number of replies, and the number of forwardings of the m^{th} microblog message, $M(t)$ is the set of microblog messages during the time period t in the microblog topic, and N_t is the total number of posted messages during the time period t in the microblog topic.

The trust between microblog users mainly considers the historical interaction between microblog users. The calculation formula is as follows:

$$T_i = (1-\lambda) + \lambda \frac{\sum_{j \in N_k} CR_i^j}{CR_i} \tag{5}$$

Where CR_i^j is the number of comments and replies from user j to user i posting the microblog message in the user set, CR_i is the total comment and reply number of microblog user i, and λ is the adjustment factor.

Based on the above two aspects of the measurement, the hybrid trust value of the microblog user is calculated as follows:

$$MT_i = \alpha GT_i + \beta T_i \tag{6}$$

Where $\alpha + \beta = 1$.

In the user set CKU_k, the top M users of the hybrid trust value are the high-quality key users that promote the dissemination and diffusion of the microblog topic. We set M equal to 5 in this paper.

4. Experimental Design and Analysis of Results

Experimental environment: The operating system is Redhat 6.5, and the database uses the ElasticSearch and Mysql databases. The algorithm implementation uses Python language. The description of experimental environment is shown in Table 1.

Table 1. The description of experimental environment

Operating system	Redhat 6.5
Database	ElasticSearch; Mysql
Programming language	Python

4.1. Data Acquisition and Preprocessing

4.1.1. Data Acquisition

During the experiment, the microblog topics in Sina platform[1] are manually labeled by the microblog topic tags. The crawler program is used to extract the corresponding topic messages and the user information according to topic features. The crawler program first crawls the microblog messages according to the topic features and the topic lifecycle time. Then, the crawler program crawls the microblog messages and the user information that forwards and comments the microblog messages. The final collected dataset includes the user and message dataset in microblog topics and the topic user historical message dataset. The framework of data acquisition is shown in Figure 1.

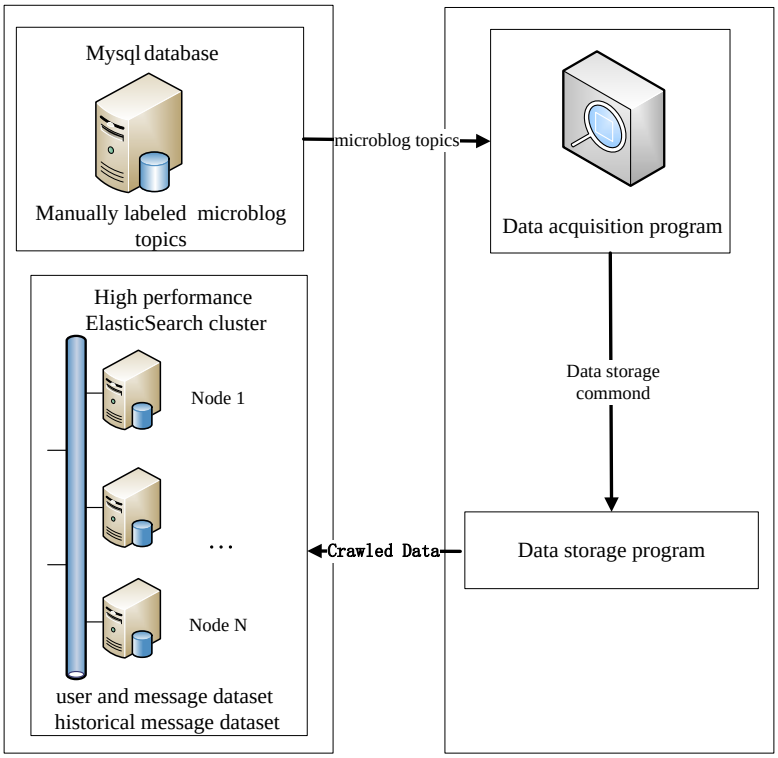

Figure 1. The framework of data acquisition

4.1.2. Data Description

The microblog topic user and message dataset contains 120,130 users and 1,404,356 microblog messages. The topic user historical message dataset contains 5,025,468 microblog messages. The microblog data obtained by the crawler has 30 fields. For the experimental study in this paper, we filter the irrelevant and repeated field fields. The final experimental reserved fields are mid, uid, parent, t, reposts_count, attitudes_count, comments_count, text,original_text, user_created_at,f ollowers_count, bi_followers_count, statuses_count, friends_count, username, verified, user_location. The descriptions of the final experimental reserved fields are shown in Table 2.

4.1.3. Data Preprocessing

The text description of a specific microblog topic generally includes the topic label text, non-topic label text, links, and other elements. In data preprocessing, the ICTClAS word segmentation system[2] developed by the Chinese Academy of Sciences is used to segment the topic tag content. Word segmentation results are cleaned up to obtain valuable descriptors to describe microblog topics. Finally, nouns and verbs are used to describe microblog topics.

---

[1] https://weibo.com/.
[2] http://ictclas.nlpir.org/.

4.2. Analysis of Experimental Results

4.2.1. Key User Mining Case Analysis

In this section, we mine the key users of the collected microblog topic dataset by using the TRank algorithm proposed in this paper, and the results are analyzed by taking the topic of "Double Eleven Soulgood Breakfast" as an example. The ranking result of key users in "Double Eleven Soulgood Breakfast" topic is shown in Table 3.

Table 2. The description of final experimental reserved fields

Fields	Description
mid	message id
uid	user id
parent	original message id
t	post time of message
reposts_count	repost count of message
attitudes_count	attitude count of message
comments_count	comment count of message
text	message text
original_text	message text of original message
user_created_at	create time of user
followers_count	follower count of user
bi_followers_count	bi_follower count of user
statuses_count	status count of user
friends_count	friend count of user
username	user name
verified	verified or not
user_location	user location

As shown in Table 3, the topic times of the top five users participating in the "Double Eleven Valley Breakfast" topic are on November 9th, November 10th, and November 11th. Analysis of the key user ranking results in the time dimension shows that the time in the topic plays a major role in key user mining. In addition, we analyze the historical microblog data of the microblog topics. In the microblog messages related to the topic, the sooner the microblog message was posted, the easier it was to be noticed by other users. The TRank algorithm proposed in this paper shows that the earlier the participating time in the topic, the more likely it was to be noticed by other microblog users, which promotes the spread of microblog topics. The analysis of the above experimental results also proves this conclusion.

The second, third, and fourth microblog users in the rankings are "Happy ZhangJiang", "Memories dedicated vest", and "Tentacle uncle". The time sequence of their participation in the topic is "Tentacle uncle", "Happy ZhangJiang", and "Memories dedicated vest". Because the key user mining method in this paper comprehensively considers the time when the user participates in the topic and the trust value of user, microblog user "Tentacle uncle" ranked lower than "Happy ZhangJiang" and "Memories dedicated vest" even if the time of participating in the topic was earlier. The TRank algorithm proposed in this paper shows that the higher the trust value of the microblog user participating in the topic, the higher the ranking of the key user of the microblog. The analysis of the above experimental results also proves this conclusion.

4.2.2. Comparison Experiments

In this section, we first select ten microblog topics in the collected data set as the microblog topic data of this experiment, including two topics that a microblog spam user participated in. Ten key users are selected for each microblog topic. Key users in microblog topics are mined by various algorithms to calculate the recognition accuracy and recall.

In this experiment, the method proposed in this paper is compared with the Twitterrank method of Weng et al. [6] and the HSA method proposed by Ye et al. [8]. The comparison results of the method are shown in Table 4.

Table 3. The ranking result in "Double Eleven Soulgood Breakfast" topic

Ranking	Username	Microblog message associated with the topic	The forwarding, comment, and like number of microblog message associated with the topic	Post time of microblog message associated with the topic
1	the daily routine of me and food	# Double Eleven Soulgood Breakfast# Erie New Products - Breakfast Ready-to-eat Cereals @Soulgood The soulgood and the yoghurt are the best together, and if you eat it every morning, will it jump?	848, 657, 914	November 9th, 19:40
2	Happy Zhangjiang	# Double Eleven Soulgood Breakfast# Erie New Products - Breakfast Ready-to-eat Cereals @Soulgood What is the best breakfast with the soulgood? I think yogurt is the best combination. Sweet and sour is love	2657, 1042, 836	November 10th 20:05
3	Memories dedicated vest	# Double Eleven Soulgood Breakfast# Erie New Products - Breakfast Ready-to-eat Cereals @Soulgood What to eat for breakfast tomorrow? It is said that the soulgood and milk are more suitable.	1810, 641, 1288	November 10th 20:30
4	Tentacle uncle	# Double Eleven Soulgood Breakfast# Erie New Products - Breakfast Ready-to-eat Cereals @Soulgood I am a tentacle uncle. What is the best breakfast with the soulgood? Do you choose milk or yogurt? Uncle will not tell you that it is yogurt~	1109, 237, 583	November 10th 18:00
5	Li Tiegen	# Double Eleven Soulgood Breakfast# Erie New Products - Breakfast Ready-to-eat Cereals @Soulgood What is the best breakfast with the soulgood? Of course, the milk is matched with the soulgood, healthy and delicious and full of nutrition! What about you, choose milk or yogurt?	428, 167, 688	November 11th 00:30

Table 4. The result of comparative experiment

Methods	Accuracy	Recall
TRank	95%	90%
Twitterrank	80%	70%
HSA	85%	75%

The comparison experiment result is shown in Table 4. By analyzing the three key users mining methods, the Twitterrank method [6] proposes an extension of PageRank algorithm to measure the influence of users, which measures the influence taking both the topical similarity between users and the link structure into account. The HSA method [8] uses the tree to interpret users' interactions in the microblog and calculate users' influence scores. A hierarchical algorithm is used for emotional analysis. These two methods do not consider the factors that promote the topic propagation and diffusion in the early stage and the trust value of the topic users, which leads to the loss of users who promote the early dissemination of the topic in the key user mining ranking results. Ranking results are sensitive to low-quality users in microblog topics. The method in this paper can mine the key users that promote the early dissemination of microblog topics and reduce the ranking of low-quality users based on the trust model, which improves the accuracy of mining high-quality key users of microblog topics.

5. Conclusions

This paper transformed the topic key user mining problem into a topic-related microblog user ranking problem. According to the topic features, a microblog message set associated with the topic was selected, and the user graph was established according to the user behavior. Time attributes were merged into the user weight calculation process. Using graph mining technology, the user trust value in the candidate key user set was evaluated to select high-quality key users.

Acknowledgements

This work was partially supported by the Doctoral Scientific Research Foundation for Henan University of Urban Construction (No. Q2017013), Key Research Development and Promotion Projects in Henan Province (Science and Technology) (No. 172102210174, 172102210105, 182402210025, 192102210258), Industry Education and Research Innovation Fund for Ministry of Education (No. 2018A01023), and Industry University Collaborative Education Project for Ministry of Education (No. 201802357006, 201802357014, 201802357032).

References

1. P. E. Brown and J. Feng, "Measuring User Influence on Twitter using Modified K-Shell Decomposition," in *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media*, pp. 18-23, Barcelona, Catalonia, Spain, July 2011
2. M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," in *Proceedings of Fourth International AAAI Conference on Weblogs and Social Media* (*ICWSM 2010*), pp. 10-17, Washington, USA, May 2010
3. Q. Fang, J. Sang, C. Xu, and Y. Rui, "Topic-Sensitive Influencer Mining in Interest-based Social Media Networks via Hypergraph Learning," *IEEE Transactions on Multimedia*, Vol. 16, No. 3, pp. 796-812, April 2014
4. C. Lee, H. Kwak, H. Park, and S. Moon, "Finding Influentials based on the Temporal Order of Information Adoption in Twitter," in *Proceedings of the 19th International Conference on World Wide Web*, pp. 1137-1138, Raleigh, North Carolina, USA, April 2010
5. 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*, pp. 1014-1022, Beijing, China, August 2012
6. J. Weng, E. P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding Topic-Sensitive Influential Twitterers," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pp. 261-270, New York, USA, February 2010
7. Y. Wu, Y. Hu, X. He, and K. Deng, "Impact of User Influence on Information Multi-Step Communication in a Micro-Blog," *Chinese Physics B*, Vol. 23, No. 6, pp. 5-12, June 2014
8. H. Ye and J. Du, "Opinion Leader Mining of Social Network Combined with Hierarchical Sentiment Analysis," in *Proceedings of 2017 Chinese Intelligent Automation Conference*, pp. 639-646, Tianjin, China, October 2017
9. H. Guo, Y. Lu, and Y. Wang, "Measuring User Influence of a Microblog based on Information Diffusion," *Journal of Shandong University*, Vol. 47, No. 5, pp. 78-83, May 2012
10. J. Li, W. Peng, and T. Li, "Social Network User Influence Sense-Making and Dynamics Prediction," *Expert Systems with Applications*, Vol. 41, No. 11, pp. 5115-5124, September 2014

Guozhong Dong, Ph.D. is a lecturer at Henan University of Urban Construction. His main research interests include data mining and information security. (20171010@hncj.edu.cn)

Bei Li, M.S. is a lecturer at Henan University of Urban Construction. Her main research interests include data mining and artificial intelligence. (85293782@qq.com)

Xinhong Wei, M.S. is an associate professor at Henan University of Urban Construction. Her main research interests include data mining and computer networks. (29813257@qq.com)

Tao Qin, M.S. (corresponding author) is a senior engineer at the National Computer Network Emergency Response Technical Team/Coordination Center of China. His main research interests include data mining and information security. (qintao@cert.org.cn)