A Method of Finding Hidden Key Users Based on Transfer Entropy in Microblog Network

Meijuan Yin[1*], Xiaonan Liu[1], Gongzhen He[1], Jing Chen[1], Ziqi Tang[1] and Bo Zhao[2,3]
[1] State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, 450001 - CN
[e-mail: raindot_ymj@163.com, nine_day@163.com,
he_gz_study@163.com,1923542221@qq.com,690245411@qq.com]
[2] Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen
Göttingen, 37073 - DE
[e-mail: bo.zhao@gwdg.de]
[3] Georg-August-Universität Göttingen
Göttingen, 37073 - DE
[e-mail: bo.zhao@gwdg.de]
*Corresponding author: Meijuan Yin

Received February 28, 2020; revised May 2, 2020; accepted June 23, 2020; published August 31, 2020

Abstract

Finding key users in microblog has been a research hotspot in recent years. There are two kinds of key users: obvious and hidden ones. Influence of the former is direct while that of the latter is indirect. Most of existing methods evaluate user's direct influence, so key users they can find usually obvious ones, and their ability to identify hidden key users is very low as hidden ones exert influence in a very covert way. Consequently, the algorithm of finding hidden key users based on topic transfer entropy, called TTE, is proposed. TTE algorithm believes that hidden key users are those normal users possessing a high covert influence on obvious ones. Firstly, obvious key users are discovered based on microblog propagation scale. Then, based on microblogs' topic similarity and time correlation, the transfer entropy from ordinary users' blogs to obvious key users is calculated and used to measure the covert influence. Finally, hidden influence degrees of ordinary users are comprehensively evaluated by combining above indicators with the influence of both ordinary users and obvious ones. We conducted experiments on Sina Weibo, and the results showed that TTE algorithm had a good ability to identify hidden key users.

Keywords: Hidden key users, microblog, obvious key users, influence, transfer entropy

1. Introduction

Weibo is one of the most popular social media in China. At present, its monthly active users exceed 400 million. It has greatly facilitated people's communication and has become a huge gathering place for public opinion. Analyzing the influence of Weibo users and finding the key users of Weibo can contribute to the supervision and guidance of public opinion in Weibo.

There are two types of key users in microblog, obvious key users and hidden key users. Obvious key users tend to have a large direct influence on ordinary users, and most of their microblogs will be explicitly forwarded, which means that when one microblog posted by an obvious key user is forwarded, the user will be automatically marked as the microblog's source by the microblog system. Hidden key users, who promote their own influence by directly affecting obvious key users, tend to have a great indirect influence on ordinary users. Obvious key users often do not directly forward the microblogs of hidden key users. Instead, they modify these microblogs slightly or copy them directly, and then post them as original microblogs. In this case, hidden key users are not marked as sources. Therefore, hidden key users have high influence and high concealment.

Kempe et al. [1] studied the "maximizing influence" problem on microblog social network. Given the parameters of the propagation model, tried to find the optimal set of seed users, which led to the largest spread scale of social influence. Yong Hua et al. [2] gave an influence maximization algorithm which found the optimal set of seed nodes with the largest propagation scale based on mixed importance. Di Shang et al. [3] assessed the influence of key influencers by analyzing the activities and structure of the social media presence of a local community.

Bashy [4] constructed forwarding cascade trees based on the topic URL, and selected static features from historical data to analyze the social influence of users. The estimated average spread scale was considered as an estimate of the user's social influence scale. Ding Zhaoyun et al. [5] believed that user forwarding behavior and the network location where users are located can promote the spread of user influence. These factors were used to measure the ability of users to disseminate information. However, the premise that these methods can accurately analyze the influence of microblog users is that the sources of microblogs are accurately marked when forwarded, which was not conducive to the identification of hidden key users.

Considering that hidden key users often promote their influence through obvious key users, we propose TTE algorithm. Obvious key users are discovered based on the user's direct influence and hidden key users are discovered indirectly through obvious key users. Firstly, the algorithm discovers the obvious key users based on the average spread scale of the microblog. Secondly, non-obvious key users are obtained from the obvious key users' follow lists. The influence of non-obvious key users on obvious key users is analyzed with their microblogs. Finally, the direct influences of two kinds of users are combined to analyze the hidden importance of non-obvious key users. In order to improve the accuracy of the influence calculation, we apply microblog topics to improve the calculation of transfer entropy [6, 7, 8]. An experiment is conducted on the Sina Weibo dataset to verify the accuracy of TTE algorithm.

The main contributions of this paper are as follows:

1.The research idea of using obvious key users as a springboard to discover hidden key users is proposed. It can provide a theoretical basis for the research of hidden key users in microblog.

2.Considering the characteristics of hidden key users, a method of discovering hidden key users is proposed. It can help discover hidden key users in microblog.

In Section 2, we give the definition of two kinds of key users and the basic idea of the proposed algorithm, TTE algorithm. Section 3 describes the steps of TTE Algorithm to find hidden key users in microblog network. Section 4 presents our experiments on Weibo dataset with the results and analysis. Finally, we conclude the whole paper in Section 5.

2. Definitions and Basic Idea

The important concept definitions of this paper are given.

Definitions 1 (obvious key users in microblog network [9]): Users with high influence but low concealment in the process of microblog propagation.

Definitions 2 (hidden key users in microblog network): Users with high influence and concealment in the process of microblog propagation.

Hidden key users spread influence by directly affecting obvious key users. Obvious key users generally propagate hidden key users' influence by forwarding the microblogs of hidden key users. During the forwarding process, the obvious key user will slightly edit and modify the microblog of the hidden key user or directly copy it without modification, and then post the microblog as original identity [10]. If an obvious key user follows an ordinary user and the obvious key user always posts microblog immediately after the ordinary user while these microblogs are highly similar, the ordinary user is likely to be a hidden key user who has a greater influence on the obvious key user.

Based on the above analysis, we propose an algorithm to find hidden key user based on topic transfer entropy, which is referred to TTE algorithm, as shown in Fig. 1. Firstly, calculate users direct influence and discover obvious key users. Secondly, the topic transfer entropy between ordinary users and obvious key users is calculated using two aspects: microblog topic similarity and time series correlation between ordinary users and obvious key users. Finally, the hidden importance will be calculated with the two kinds of direct influence.

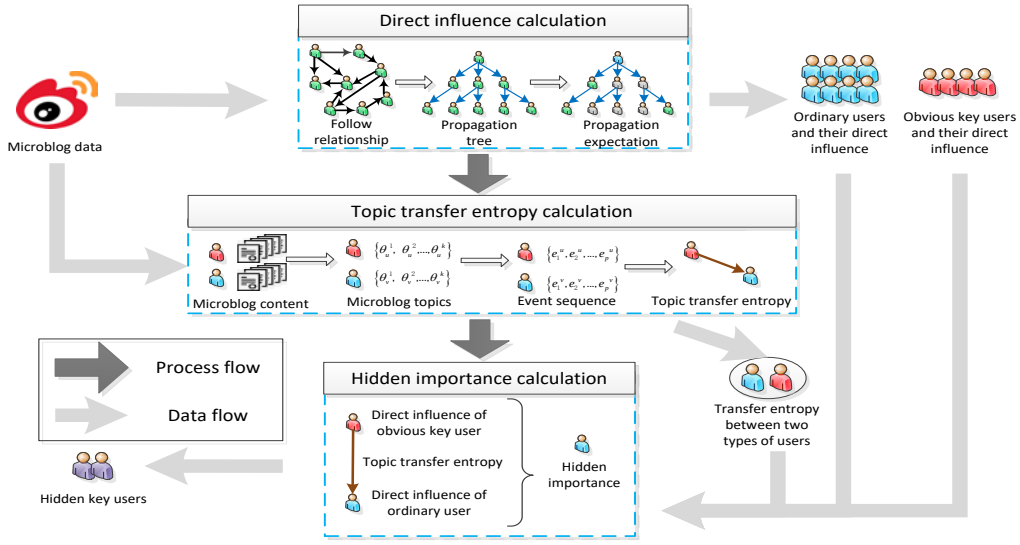


Fig. 1. Steps of TTE Algorithm

In summary, this article focuses on the following issues:

1. How to find obvious key users.
2. How to get the user's microblog topic and calculate the topic similarity between different microblogs.
3. How to calculate the topic transfer entropy of ordinary user to obvious key user.
4. How to comprehensively assess the hidden importance of ordinary users.
5. How to assess the effectiveness of hidden key user discovery results.

3. The Algorithm for Finding Hidden Key Users in Microblog Network

3.1 Finding Obvious Key Users

Obvious key users have a high direct influence on other users in the microblog network. Most of the current methods measure the direct influence of users using the spreading scale of microblogs posted by users. We use the method of the literature [11] to measure the user's direct influence. We calculate the propagation expectation based on a microblog propagation tree which reflects the following relationship between users. We can discover the obvious key user with this indicator.

The calculation method of user u's microblog propagation expectation is shown in formula (1).

$$Spreadrange_u = \sum_{v \in follower_u} \left(Probability_{v \to u} * Spreadrange_v \right) \tag{1}$$

where $Spreadrange_u$ indicates the user's microblog propagation expectation, $follow_u$ indicates the direct fan user set of u in the propagation tree, v indicates a member of $follow_u$, $Probability_{v \to u}$ indicates the forwarding probability of v to u's microblog.

3.2 Extraction of Microblog Topic Event Sequence

An event consists of content and time, and a user's microblog topic event includes posting time and content of the user's microblog for each topic. The user's microblog event sequence, which is the premise of calculating the transfer entropy, consists of time-ordered microblog events. The extraction method of the user microblog event sequence will be described in detail.

3.2.1 Symbol Definition

$\{w_1, w_2, ..., w_n\}$ indicates all microblog users where n indicates the total number of users, and w_i indicates the set of all microblogs posted by the i-th user during the observation time interval. $\{m_1, m_2, ..., m_k\}$ indicates microblog set of each user where m_j indicates the j-th microblog posted by the user during the observation time interval, and k indicates the total number of user microblogs. Each microblog is described as {c, p, t}, where c represents microblog text content, p represents microblog topic, and t represents the posting time of microblog.

3.2.2 Extraction of Microblog Event Topics

If we directly use the text of microblog to express the content of microblog, it will make the analysis more difficult and it is easier to cause errors. Therefore, we extract the topic of microblog with LDA model [12, 13] and use it to represent the content of microblog. The LDA model, including texts, topics, and words, is a three-layer Bayesian model, which is a common method for topic extraction. The basic idea is that each microblog text c can be repre-

sented as a multi-distribution of k topics, and each topic can be represented as a multi-distribution of all words in a vocabulary. LDA three-layer Bayesian network is shown as follows.

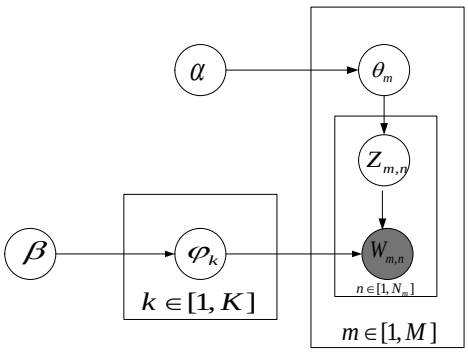


Fig. 2. LDA three-layer Bayesian network

θ indicates microblog-topic distribution vector. $\theta=\{\theta^1,\theta^2,\mathrm{K},\theta^k\}$, where $\theta^1,\theta^2,\mathrm{K},\theta^k$ represents k-th element in the microblog topic distribution vector, which is the probability distribution of text c in k topics.

3.2.3 Extraction of Microblog Event Sequence

The time interval for the transfer entropy event sequence is fixed while the time interval for users to post microblog is not fixed. We need to correct the release timing of the user's microblog and adopt a flexible time interval. The correction method is as follows.

For all microblogs a user posted in an observation time interval T, we use LDA to extract the topic of each microblog. Then we apply the hierarchical clustering algorithm [14, 15, 16] to cluster the microblogs in the time interval. When the similarity of the topic vectors of two adjacent microblogs is greater than a certain threshold, these two microblogs will be put into one cluster. Multiple microblog clusters will be obtained in an observation time interval. The clustering results are shown in **Fig. 3**.

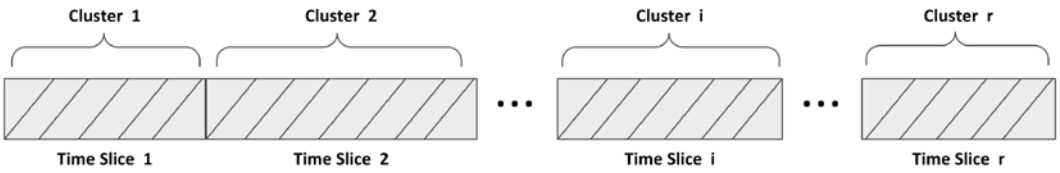


Fig. 3. Microblog clustering

Finally, each microblog cluster of a user in the observation time interval is regarded as one microblog. The release time and topic of the microblogs are represented by the time and topic of the first posted microblog in the cluster. Thus, the sequence of microblog events of the user e = {e_1, e_2, ..., e_r} in the observation time interval T is obtained. The topic vector similarity calculation method is shown in formula (2).

$$TopicSim_{n,n+1}=\frac{\sum_{i=1}^{k}\left(\theta_n^i\times\theta_{n+1}^i\right)}{\sqrt{\sum_{i=1}^{k}\left(\theta_n^i\right)^2}\times\sqrt{\sum_{i=1}^{k}\left(\theta_{n+1}^i\right)^2}} \tag{2}$$

3.3 The Calculation of Microblog Topic Transfer Entropy

3.3.1 Traditional Transfer Entropy

Transfer entropy is an effective tool to analyze the interaction between two systems. When two systems generate a series of events on a sequence of time, the degree of influence of one system event on another system event can be calculated by the transfer entropy between two event sequences.

Suppose there are two systems, A and B. The formalized description of the calculation method of the transfer entropy from System A to System B is as follows:

The sequence of events that occur in System A is {x1, x2, ... xl}, the size of the set is l. The sequence of events that occur in System B is {y1, y2, ... yl}. $P\left(x_{n+1} \mid x_n, y_n\right)$ indicates the probability that x_{n+1} will occur during the $n+1$-th time period under the condition that events x_n and y_n occurred during the n-th time period. $P\left(x_{n+1} \mid x_n\right)$ indicates the probability that event x_{n+1} occurs during time period n+1 under the condition that event x_n occurred during time period n. If $P\left(x_{n+1} \mid x_n, y_n\right) > P\left(x_{n+1} \mid x_n\right)$, it infers that System B has an influence on System A [17]. Transfer entropy measures the magnitude of this effect and is calculated as follows.

$$h_1 = -\sum_{x_{n+1}} P\left(x_{n+1}, x_n, y_n\right) log_a P\left(x_{n+1} | x_n, y_n\right) \tag{3}$$

$$h_2 = -\sum_{x_{n+1}} P\left(x_{n+1}, x_n, y_n\right) log_a P\left(x_{n+1} | x_n\right) \tag{4}$$

$$TE_{B\to A} = h_2 - h_1 = -\sum_{x_{n+1}, x_n, y_n} P\left(x_{n+1}, x_n, y_n\right) log_a \left(\frac{P\left(x_{n+1} | x_n, y_n\right)}{P\left(x_{n+1} | x_n\right)}\right) \tag{5}$$

If the value of the transfer entropy is greater than zero, it indicates that System B may have an effect on System A, and the larger the value is, the greater the influence is.

3.3.2 Microblog Topic Transfer Entropy

If user A posts microblogs on Weibo following user B and the topics of their microblogs are similar, it indicates that the A's microblogs follow B's in both time and content. We can infer that B has a certain influence on A. Each user in Weibo can be regarded as a microblog posting system, and one microblog posting behavior of the user corresponds to one event in the system. The transfer entropy of the two users' microblog posting behavior, namely the microblog topic transfer entropy, can measure the degree of follow-up of one user to another from both time and content. Calculating the microblog topic transfer entropy requires the user's microblog event sequence, and the user's microblog event sequence is described below.

After clustering microblogs posted by user S, we obtain the event sequence $\{e_1, e_2, ..., e_p\}$, where p is the length of the sequence of events. $e_i \in \{0,1\}\ (1 \le i \le p)$. e_i=1 indicates that the user posted microblog on the i-th time period and e_i=0 indicates that user S has not posted microblog during this time period. The user's events can be represented as a sequence of 0, 1 in a sequence of time, which is the basis of the calculation of transfer entropy.

$\left\{e_1^u, e_2^u, ..., e_p^u\right\}$ indicates event set of obvious key user u, and $\left\{\theta_1^u, \theta_2^u, ..., \theta_p^u\right\}$ indicates its microblog topic sequence. $\left\{e_1^v, e_2^v, ..., e_p^v\right\}$ indicates event set of v who is a general user followed by u, and $\left\{\theta_1^v, \theta_2^v, ..., \theta_p^v\right\}$ indicates its microblog topic sequence.

3.3.3 Calculation of Microblog Transfer Entropy

Only the time of the events of the two systems is considered when the traditional transfer entropy infers the interaction between two systems. Applying the microblog topic similarity to improve the traditional transfer entropy can more accurately measure the influence of one microblog user on another user. Therefore, the calculation method of the microblog topic transfer entropy is shown in formula (6).

$$TE_{v\to u} = -\sum_{e_{n+1}^{u},e_n^{u},e_n^{v}} \left(P\left(e_{n+1}^{u},e_n^{u},e_n^{v}\right)\log_a\left(\frac{P\left(e_{n+1}^{u}|e_n^{u},e_n^{v}\right)*Topicsim\left(\theta_{n+1}^{u},\theta_n^{v}\right)}{P\left(e_{n+1}^{u}|e_n^{u}\right)}\right)\right) \tag{6}$$

where $Topicsim\left(\theta_{n+1}^{u},\theta_n^{v}\right)$ indicates the cosine similarity of the first microblog posted by the user u in the n+1th time period and the first microblog posted by the user v in the nth time period. The calculation method of cosine similarity is shown in formula (2).

3.4 The Calculation of Hidden Importance

In this paper, we use hidden importance to reflect the possibility that a user is a hidden key user. The indicator can be measured by three sub-indicators:

1. The Direct Influence of the User. The direct influence of hidden key users must not be high, and the lower the direct influence, the higher the possibility that it is hidden key users. Therefore, the user's hidden importance is inversely proportional to the direct influence.

2. microblog Topic Transfer Entropy. The higher the user's influence on the obvious key users, the greater the transfer entropy of the microblog topic, which can reflect the user's indirect influence. Therefore, the hidden importance is proportional to the microblog topic transfer entropy.

3. Direct Influence of Obvious key users. According to the previous analysis, the hidden importance is directly proportional to the direct influence of obvious key users.

In summary, the calculation method of hidden importance is shown in formula (7):

$$HiddenInf_v = \frac{1}{Influence_v} * \sum_{u\in KeyFAN_v} TE_{v\to u} \times Influence_u \tag{7}$$

where $HiddenInf_v$ represents the hidden importance of the user v, and $KeyFAN_v$ represents an obvious key user in that user's fan set. It can be seen from the formula that the user's hidden importance is inversely proportional to its direct influence, proportional to the direct influence of the obvious key users affected by it, and is proportional to the microblog topic transfer entropy.

4. Experiment and Analysis

In this section, we verify the validity of the proposed TTE algorithm. Firstly, we give the Weibo datasets and its processing method. Then two important parameters of TTE algorithm are determined by some experiments. And on this basis, we find hidden key users in the dataset by using TTE algorithm and analyze its accuracy.

4.1 Datasets

4.1.1 Weibo dataset

We used real Weibo dataset to verify the validity of the algorithm. The dataset spans from March 2012 to March 2014 with a total of 3 million Weibo users. It covers hot topics of so-

cial and national public opinion such as the "Yaan earthquake", "the death of Linwu hawkers" and "the dispute over the Diaoyu Islands". We can use this dataset to mine the Weibo users who are actively spreading negative information during the hot topic communication process.

4.1.2 Data Preprocessing

The number of posts in Sina Weibo is huge [18], there is a lot of noise information such as zombie users, advertising users and garbage microblogs in the Sina Weibo platform.

These elements are abundant in the experimental data. In order to facilitate the analysis of the data by the algorithm and improve the efficiency and accuracy of the experiment, it is necessary to clean the original data of the microblog.

After statistics, zombie users generally do not post microblog. The number of microblogs posted by most Weibo users within one day is no more than 3, so we believed that the Weibo account which posts more than 5 microblogs is generally an advertising account microblogs, which are not forwarded, liked or commented, are of little value on the research of user influence, and they are not accounted in the experiment.

According to the above analysis, if a Weibo account or a microblog does not satisfy any of the following characteristics, it was filtered out.

1. Each Weibo account posts at least one Weibo per week;
2. A Weibo account does not post more than 5 microblogs per day;
3. A Weibo has been forwarded, liked or commented at least once.

4.2 Parameter Selection Experiment

4.2.1 Topic Similarity Threshold

Firstly, we manually select 10,000 pairs of microblogs with the same topic and 10,000 microblogs with different topics. Then the topic similarity of the two groups of microblogs is calculated separately with the method of topic similarity calculation. Finally, the distribution of similarity between the two groups is statistically analyzed. The statistical results are shown in **Fig. 4**.

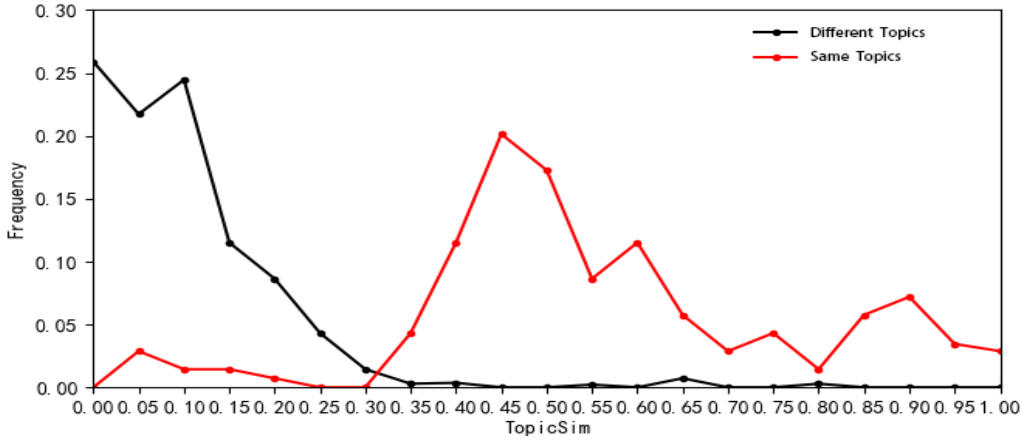


Fig. 4. Topic Similarity Distribution of Two Groups of Weibo

In **Fig. 4**, the topic similarity of the microblogs with the same topic is mostly above 0.33, and the topic similarity of the microblogs with different topics is mostly less than 0.33. So, we set the topic similarity threshold of microblog combination to 0.33.

4.2.2 Selection of Obvious Key User

Obvious key users are the foundation of mining hidden users, therefore the policy to accurately select obvious key users according to their direct influence is an important parameter for TTE algorithm. We determine the policy by analyzing the distribution of Weibo users' direct influence values.

We evaluated the direct influence of 3 million users by the method in section 3.1 and conducted a statistical analysis of the direct influence values. The statistical result of the direct influence value distribution is as shown in **Fig. 5**.

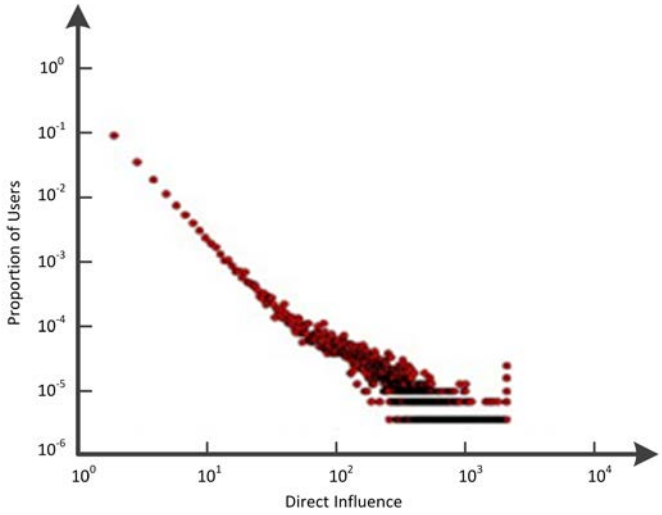


Fig. 5. Topic Similarity Distribution of Two Groups of Weibo

In **Fig. 5**, the values of 3 million user's influence are from 0 to 10000 and roughly obey the power-law distribution, and most users' influence is found to be in the interval of (0,100). About 90% of users have direct influence within 100 and the number of users with influence greater than 100 is dropped dramatically. It follows that the influence of users with the top 10% influence value is significant and these uses can be taken as the candidate set of obvious key users.

Experts in the field of Weibo security pointed out that the direct influence of hidden key users is closer to those ordinary users with the later 90% influence value, and the direct influence of obvious key users is usually very high as influenced by the hidden key users. So, we only take the top 30% of the candidate obvious key users as the real obvious key users, that is the users with the top 0.3% influence value.

4.3 Hidden Key Users Finding Experiment

4.3.1 Evaluation Method of Experimental Results

As there are no acknowledged labelled dataset and evaluation criterion for hidden key users and no comparison methods of finding hidden key users at present, we use expert scoring to label the relative hidden importance of users, and evaluate the accuracy of suspected hidden key users discovered by TTE algorithm based on the expert scoring results. The criterion of expert scoring is shown in **Table 1**.

Table 1. Criterion for expert scoring

Score of hidden importance	Possibility of being a hidden key user
1-2	unlikely
3-4	low possibility
5-6	normal possibility
7-8	high possibility
9-10	very likely

As shown in **Table 1**, the experts score the hidden importance of Weibo users on a scale of 1 to 10, and the higher the user's score, the more likely it is that the user is a hidden key user. In order to reduce the deviation of scoring results caused by different understanding of the criterion for different experts, 10 scores are divided into 5 intervals with 2 scores in each interval for the expert to choose, which could improve the consistency of expert scoring result to a certain extent.

The criterion of TTE algorithm to calculate the hidden importance of users is different with the one of expert scoring. Therefore, the accuracy of TTE algorithm cannot be directly evaluated by comparing the value of TTE algorithm with that of expert scoring for the same user. But the higher the value of hidden importance given by both methods, the greater the possibility of the user being a hidden key user is. Therefore, users can be sorted in descending order respectively according to the hidden importance obtained by two methods, and both sequences represent the descending order of relative hidden importance of users. Thus, the higher the consistency of two user sequences, the more accurate the hidden importance calculated by TTE algorithm is.

In this paper, Pearson correlation coefficient [19] is chosen to measure the consistency of two user sequences. Pearson correlation coefficient is a common indicator to measure the correlation degree of two sorts. The larger the Pearson coefficient, the more correlated two sorts are, that is, the smaller the difference between two sequences is. The corresponding relationship between the value of Pearson coefficient and the correlation degree is shown in **Table 2**.

Table 2. Pearson coefficient and correlation degree table

Value of Pearson Coefficient	Correlation Degree
0.8-1.00	extreme
0.6-0.79	high
0.4-0.59	moderate
0.2-0.39	low
0.0-0.19	very low

4.3.2 Experimental Results

At first, according to the direct influence of 3 million Weibo users in the dataset, the top 0.3%, that is 9000 users, were selected as the obvious key users. Then we extracted users followed by these obvious key users, filtered out the obvious key users among these users and obtained the candidate hidden key user set, which contains 49,536 users. Finally, the hidden importance of each user in the candidate hidden key user set was calculated by TTE algorithm. The value range of the hidden importance obtained is (0, 9). The user's hidden importance drops obviously and only the top 20 users' hidden importance is greater than or

close to 3 when we sorted all candidate hidden key users by the hidden importance in descending order. **Table 3** shows the ID of the top 20 users and their hidden importance.

Due to the scarcity of hidden key users and the limitation on data scale of manually labelling dataset, we only selected the top 20 users with high hidden importance as suspected hidden key users discovered by TTE algorithm to evaluate the accuracy of TTE algorithm.

In order to further reduce the deviation of expert scoring results, three experts in microblog security field were invited to score the hidden importance of these 20 suspected hidden key users according to the criterion in **Table 1**, and the average score of three experts was taken as the final score of hidden importance of each user. The average score of three experts for each user is shown in the last column in **Table 3**.

Table 3. Top 20 suspected hidden key users

ID (Original Weibo ID)	Hidden Importance	Direct Influence Ranking	Average Score by Experts
天王黄金archer	8.93	1025678	9.32
三峡西北望	8.64	2620554	7.18
木木夕沥漓	8.32	1370945	6.43
胜永王	7.96	893706	6.09
毛小秀Nico437	7.7	564087	8.84
励步儿童教育_昆明	6.46	335712	8.46
张兴军	5.71	450089	9.13
杨小花yy	5.62	1102364	3.25
3045226700nZL	5.58	904355	6.66
机智的夏尔	5.45	1066528	5.73
ahh_ching	5.43	2706354	5.59
可靠的霍洛霍洛1990	5.41	1856972	5.12
陆勤医生	4.95	1212457	6.73
杉树林V	4.65	545879	4.51
IP用户通	4.29	1130465	7.48
futuregardenCAFE	3.95	1387966	4.44
黄蔓崇	3.66	4067563	3.81
笨小孩047	3.43	978459	4.63
宝宝舞夏	3.25	334678	3.45
在路上jeffrey	2.92	382699	3.11

We first analyzed the crypticity of 20 suspected hidden key users in **Table 3**.

The direct influence rankings of these users are listed in the third column in **Table 3**. It can be seen that the top one in the 20 users ranks 334678, which is 11.2% of 3 million and not in the top 10% range, and 17 users (85% of 20 users) are beyond the top 15% range, and 14 users (70% of 20 users) are beyond the top 30% range. Through further analysis of the data, we found that Weibo of these hidden key users is not marked as the source in Weibo dissemination and thus the direct influence of these users is not high, which further illustrates the crypticity of hidden key users.

Then we analyzed the expert scoring results of 20 suspected hidden key users with their hidden importance obtained by TTE algorithm in **Table 3**.

Weibo security experts believe that users with scores about or above 4.5 have further research value and can be regarded as key users with crypticity. Among the top 20 suspected hidden key users discovered by TTE algorithm, about 80% of them have an average expert score more than 4.5, which reflects the accuracy of TTE algorithm.

We then calculated the Pearson coefficient between the user ranking by the first column (that is the user ranking by hidden importance of TTE algorithm) and the user ranking by expert average scores in the fourth column in descending order for the 20 suspected hidden key users. The Pearson correlation coefficient of two user rankings is 0.71. The value shows that two sorts are highly correlated by **Table 2**, which further proves the accuracy of TTE.

Comparing the difference between the average score of the experts and the calculated hidden importance, it is found that there are some special hidden key users. For example, the user whose ID is "张兴军" in **Table 3**. which is believed highly likely to be a hidden key user by experts. According to experts' scores, its hidden importance should be ranked the second, but the hidden importance of the user calculated by TTE algorithm ranks the seventh. Further analysis found that the user's microblog is very short, and the affected microblogs posted by the obvious key users generally added a lot of other content, which makes it difficult to accurately calculate the topic similarity of two users' microblogs and causes some certain errors.

5. Conclusion and Future Work

In this paper, we innovatively propose the concept of hidden key users, apply microblog topic to improve the traditional transfer entropy calculation method and propose TTE algorithm to evaluate user's hidden importance. A lot of experiments were conducted on Sina Weibo's dataset to verify the effectiveness of TTE algorithm.

We will explore the following directions in the future:

1. How to set the topic similarity threshold for controlling microblog combination, which is a key issue.

2. We will focus on improving the similarity calculation of the microblog topics. Due to the flexibility of Chinese and the short text characteristics of microblog, microblogs on the same topic may have high semantic similarity, but there are huge differences in literal expression, which may lead to difficulty in calculating the similarity of microblog topics.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant no. U1636219) and the National Key Research and Development Program of China (Grant no.2016YFB0801303, 2016QY01W0105).

References

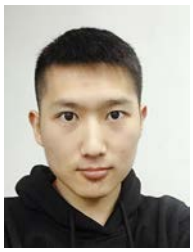
[1] D. Kempe, J. Kleinberg and Tardos éva, “Maximizing the spread of influence through a social network,” in *Proc. of KDD '03 Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 137–146, 2003. Article (CrossRef Link)

[2] Yong Hua, Bolun Chen, Yan Yuan, Guochang Zhu and Jialin Ma, “An Influence Maximization Algorithm Based on the Mixed Importance of Nodes,” *Computers, Materials & Continua*, vol. 59, no. 2, pp. 517-531, 2019. Article (CrossRef Link)

[3] Di Shang and Mohammed Ghriga, "Examining the Impacts of Key Influencers on Community Development," *Computers, Materials & Continua*, Vol. 61, No. 1, pp.1-10, 2019. Article (CrossRef Link)
[4] E. Bakshy E, J. M. Hofman, W. A. Mason and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proc. of WSDM '11 Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 65–74, 2011. Article (CrossRef Link)
[5] Z. Y. Ding, Y. Jia, B. Zhou, et al, "Measuring the spreadability of users in microblogs," *Journal of Zhejiang University (English Edition) (C Series: Computer and Electronics)*, vol. 14, no. 9, pp. 701–710, 2013. Article (CrossRef Link)
[6] T. Bossomaier, L. Barnett, M. Harré, J. T. Lizier, "Transfer Entropy," *An Introduction to Transfer Entropy*, pp. 65–95, 2016. Article (CrossRef Link)
[7] L. Barnett, A. B. Barnett and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variablesc," *Physical Review Letters*, vol. 103, no. 23, pp. 238701-238701, 2009. Article (CrossRef Link)
[8] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, pp. 461–464, 2000. Article (CrossRef Link)
[9] Zheng Yongguang, Yue Kun, Yin Zidu and Zhang Xuejie, "Efficient key user selection method in large-scale social networks," *Journal of Computer Applications*, vol. 37, no. 11, pp. 3101-3106, 2017. Article (CrossRef Link)
[10] H. He, Z. Yu, B. Guo, X. Lu and J. Tian, "Tree-Based Mining for Discovering Patterns of Reposting Behavior in microblog," in *Proc. of ADMA 2013. Lecture Notes in Computer Science*, pp. 372–384, 2013. Article (CrossRef Link)
[11] Ziqi Tang, Meijuan Yin and Junyong Luo, "Disseminating Quality-Based Analysis of microblog Users' Influencing Ability," in *Proc. of the 4th International Conference on Cloud Computing and Security*, pp. 499–514, June 8-10, 2018. Article (CrossRef Link)
[12] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, no. 3, pp. 993–1022, 2003. Article (CrossRef Link)
[13] T. L. Griffiths, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. 1, pp. 5228–5235, 2004. Article (CrossRef Link)
[14] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241-254, 1967. Article (CrossRef Link)
[15] P. P. Rodrigues, J. Gama, and J. P. Pedroso, "Hierarchical Clustering of Time-Series Data Streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 5, pp. 615–627, 2008. Article (CrossRef Link)
[16] Bo Xiao, Zhen Wang, Qi Liu and Xiaodong Liu, "SMK-means: An Improved Mini Batch K-means Algorithm Based on Mapreduce with Big Data," *Computers, Materials & Continua*, vol. 56, no. 3, pp. 365–379, 2018. Article (CrossRef Link)
[17] Seth A, "Granger causality," *Scholarpedia*, vol. 2, no. 7, pp. 1667-1667, 2007. Article (CrossRef Link)
[18] Kai Dong, Taolin Guo, Xiaolin Fang, Zhen Ling and Haibo Ye, "Estimating the Number of Posts in Sina Weibo," *Computers, Materials & Continua*, vol. 58, no. 1, pp. 197–213, 2019. Article (CrossRef Link)
[19] I. Cohen, J. Chen, Y. Huang et al., "Pearson Correlation Coefficient," *Noise Reduction in Speech Processing*, vol. 2, pp. 1-4, 2009. Article (CrossRef Link)

Meijuan Yin was born in Anhui Province, China at November, 1977. She was conferred a M.S. in computer science by State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou, China, in 2003. She is working on the Ph.D. in computer software and academic of the same Laboratory. After graduating from the Laboratory, she became an assistant of the same Laboratory in 2003 and turned to a lecturer in 2005. Her current research interests include data mining, social network analysis, and information security.

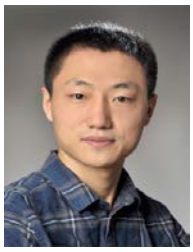
Xiaonan Liu was born in Liaoning Province, China. He was conferred a M.S. in computer science by State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou, China, in 2006. He is working on the Ph.D. in computer software and academic of the same university. He graduated from the university, and became an assistant in 2000 and turned to a lecturer in 2006. Now, he is an associate professor of State Key Laboratory of Mathematics Engineering and Advanced Computing of the university. His research interests include binary translation, compile, and decompile.



Gongzhen He was born in Nanyang, Henan, China in 1991. He received the B.S. degrees in information management from the Nanjing University of Science and Technology at Nanjing, China, in 2012.He is currently pursuing the M.S. degree in computer technology in the State Key Laboratory of Mathematical Engineering and Advanced Computing. His research interests include data mining, social network analysis, and information security.

Jing Chen was was born Jiaozuo, Henan Province, China in 1990. She received a master's degree in computer science and technology from the State Key Laboratory of Mathematical Engineering and Advanced Computing at Zhengzhou, China, in 2016. From 2016 to the present, she is an assistant professor of the Department of Big Data Analysis at the State Key Laboratory of Mathematical Engineering and Advanced Computing. Her research interests include data mining, natural language processing, and network data analysis.

Ziqi Tang was born in TianFeng Village, Jiang Xi province, China in 1994. He received the B.S. and M.S. degrees in Information Engineering University, ZhengZhou, in 2018. From 2014 to 2018, he was a Research Assistant with the State Key Laboratory of Mathematical Engineering and Advanced Computing. His research interest is social media data analysis. He holds two patents



Bo Zhao received his B.S. degree in Computer Science and Technology and M.S. degree in Computer software and Theory from Tsinghua University, Beijing, China in 2011 and National Digital Switching System Engineering & Technological Research Center (NDSC), Zhengzhou, Henan, China in 2014, respectively. His main Research work before is dependency analysis and parallelism exploitation for applications in High Performance Computing. He has been a doctoral candidate of NDSC from 2014 and his research topic is Construction for unified vectorization framework. Currently, he is studying for his doctoral degree at GWDG. His current research interests are consumer behavior analysis and prediction based on big data, recommendation systems and social network analysis.