BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 03/06 pp34–46
Volume 1, Number 1, March 2018
DOI: 10.26599/BDMA.2018.9020004

Event Detection and Identification of Influential Spreaders in Social Media Data Streams

Leilei Shi, Yan Wu, Lu Liu*, Xiang Sun, and Liang Jiang

Abstract: Microblogging, a popular social media service platform, has become a new information channel for users to receive and exchange the most up-to-date information on current events. Consequently, it is a crucial platform for detecting newly emerging events and for identifying influential spreaders who have the potential to actively disseminate knowledge about events through microblogs. However, traditional event detection models require human intervention to detect the number of topics to be explored, which significantly reduces the efficiency and accuracy of event detection. In addition, most existing methods focus only on event detection and are unable to identify either influential spreaders or key event-related posts, thus making it challenging to track momentous events in a timely manner. To address these problems, we propose a Hypertext-Induced Topic Search (HITS) based Topic-Decision method (TD-HITS), and a Latent Dirichlet Allocation (LDA) based Three-Step model (TS-LDA). TD-HITS can automatically detect the number of topics as well as identify associated key posts in a large number of posts. TS-LDA can identify influential spreaders of hot event topics based on both post and user information. The experimental results, using a Twitter dataset, demonstrate the effectiveness of our proposed methods for both detecting events and identifying influential spreaders.

Key words: event detection; microblogging; Hypertext-Induced Topic Search (HITS); Latent Dirichlet Allocation (LDA); identification of influential spreader

1 Introduction

Along with the increasing popularity of social networking in recent years, microblogging services, as a social media platform, have also been developing and attracting users at a rapid pace[1–3]. The huge volume of data generated for the most important events in microblogging requires the determination of hot events as well as the identification of key posts related to these events and the influential spreaders with the potential to help others track these hot events. Therefore, it is crucial that hot events be detected in microblogs.

Recently, event detection methods based on topic models have been increasing in popularity[4,5]. For example, Probabilistic Latent Semantic Analysis (PLSA)[6] and Latent Dirichlet Allocation (LDA)[7] are two important approaches for detecting hidden variables in microblogs. These methods model word occurrences based on probabilistic theory and measure the topical similarity among words. Although researchers have made significant efforts to detect target events in social networks based on a single source, in a crisis we often want to analyze key event-related posts contributed by different social users. Thus far, scant attention has been paid to the problem of detecting events from among the integrated and ambiguous views contributed by different users.

• Leilei Shi, Yan Wu, Lu Liu, Xiang Sun, and Liang Jiang are with School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang 212013, China.
• Lu Liu is also with Department of Computing and Mathematics, University of Derby, UK. E-mail: l.liu@derby.ac.uk.
∗ To whom correspondence should be addressed.
Manuscript received: 2017-09-09; accepted: 2017-11-29

Additionally, existing event detection models[6–11] require human intervention to detect the number of topics, which greatly reduces the efficiency and accuracy of event detection. Furthermore, most existing methods focus only on event detection and fail to investigate the key posts or influential spreaders who play an important role in the dissemination of critical events. This makes it difficult for Internet watch officers to track critical events in a timely manner.

To address this failing, in this paper, we propose a Hypertext-Induced Topic Search (HITS) based[12,13] Topic-Decision method (TD-HITS) that can automatically detect the number of topics and identify key posts from among a large number of posts. We also propose a Latent Dirichlet Allocation (LDA) based Three-Step model (TS-LDA), which can identify the most influential spreaders of hot events based on both post and user information. Using a Twitter dataset for our study, our experimental results demonstrate the effectiveness of our proposed methods for both event detection and the identification of influential spreaders.

The main contributions of this paper are as follows:

- We propose an HITS-based topic-decision method. This approach creates a smaller high-quality training dataset by selecting high-quality posts and influential users from among a collection of users and posts, which largely reduces the impact of irrelevant posts and ordinary users, and improves the efficiency and accuracy of event detection compared with those of existing methods[4,5,8–11]. Moreover, the proposed approach can automatically detect the number of topics and identify key event-related posts from among a large number of posts, which further improves the efficiency and accuracy of event detection and outperforms existing methods[8–11,14–18].
- We propose an LDA-based three-step model that detects critical events based on the number of topics and identifies influential spreaders involved in sharing these critical events. This model utilizes both post and user information, which can improve our understanding of who is involved in these critical incidents.
- We conducted experiments to evaluate the performance of our proposed models. The experimental results on a Twitter dataset demonstrate the efficiency and accuracy of our models in event detection and the identification of influential spreaders.

The rest of this paper is organized as follows. In Section 2, we introduce previous studies of event detection. In Section 3, we describe the proposed TD-HITS method. We introduce the TS-LDA model in Section 4 and discuss the experimental analysis and the obtained results in Section 5. In Section 6, we draw our conclusions.

2 Related Work

In recent years, event detection has been the focus of a wide range of research, especially from the social media perspective, due to its openness and data availability (e.g., Twitter access through Twitter API[19] and Facebook access through Facebook API[20]). Existing event detection models for social media are categorized as either feature-pivot[14] or document-pivot[15] models.

Feature-pivot models are used to study the distributions of words and to detect events by grouping words together. For example, Mathioudakis and Koudas[16] detected events by grouping bursty words. However, this method does not have a robust probabilistic foundation and focuses only on event detection; it fails to identify key event-related posts or the influential spreaders involved in these critical incidents. Wavelet analysis[21] has been applied to the frequency-based raw signals of words in building signals for individual words, and filters trivial words by examining their corresponding signal auto-correlations. This method detects events using a modularity-based graph partitioning technique. However, it also focuses only on event detection and does not take into account key posts or influential spreaders, which increases the complexities involved in promptly tracking and controlling events.

Document-pivot models detect events by clustering documents according to the semantic distances between them. For example, Wang et al.[9] proposed a pLSA-based model that exploits this document-pivoting concept to find correlated bursty patterns across multiple text streams. Alsumait et al.[10] proposed the use of an LDA topic model to model the topics in text streams. The authors also used an evolution matrix to record the varieties of topics, which achieved good performance. Diao et al.[8] proposed an LDA-based model that exploits the same pivoting idea to identify bursty global events. Li et al.[11] proposed a Bursty Event dEtection (BEE) topic model that detects new bursty events by modeling these events. However,

these event detection methods all show vulnerabilities in the automatic detection of the number of topics and the identification of associated key posts and influential spreaders involved in these critical events. Document-pivot-based event detection models have long been applied in Topic Detection and Tracking (TDT) programs[17, 18]. TDT systems provide general outlines and fundamentals regarding event detection. However, noisy posts and the great numbers of ordinary users make these methods unsuitable for either critical event detection or the identification of influential spreaders for large quantities of social media data.

In summary, the above models do not tend to perform well in event detection in the following respects: First, the characteristics of microblogging, such as the relationships between users and their posts, cannot effectively address the influence of users and the importance of posts. These methods[16, 21] are focused only on event detection by the grouping of words. Second, existing methods[6–11] consider only event detection and are not concerned with the discovery of key posts. Finally, no attempt is made to identify influential spreaders related to hot events and most existing methods[8–11, 14–18] involve the manual subjective selection of the number of topics.

To tackle the problems outlined above, in this paper, we propose the TD-HITS method, which can automatically detect the number of topics and identify key posts from among a large number of noisy posts. Based on the TD-HITS model, we further propose the TS-LDA model, which is a document-pivot model. In the proposed TD-HITS model, noisy posts and ordinary users are effectively removed from the selection, and with the proposed TS-LDA model, there is no need to set up the number of topics manually in advance as it effectively detects hot events and identifies influential spreaders. Finally, our proposed methods exhibit better efficiency and accuracy in event detection and the identification of influential spreaders by addressing the above-noted drawbacks of existing methods[8–11, 14–18, 21].

3 TD-HITS Method

The TD-HITS method has two modules: first, the HITS algorithm is used to create a smaller high-quality training data set by extracting high-quality posts and influential users from the large pool of posts and users. Second, a topic-decision method is used to automatically detect the number of topics and to discover key posts from among a large number of posts.

Figure 1 illustrates the TD-HITS process involved in event detection and the identification of key posts.

3.1 Extracting high-quality posts and influential users with the HITS algorithm

3.1.1 Extracting high-quality posts

In the original HITS method[12], a link is used to represent the hyperlinks between web pages. In our TD-HITS method, however, a link represents an operational relationship between a user and a post, such as publishing or commenting. For example, given a post in an undirected network $G = (V, E)$, where $V = \{V_1, V_2, \ldots, V_n\}$ is a set of n posts, and E is a set of undirected edges between posts, the n posts and their connections are interpreted by an adjacency matrix: $\boldsymbol{A} = [A_{ij}]_{n\times n}$, if post V_i and V_j are connected, then $A_{ij} = 1$, otherwise, $A_{ij} = 0$. The user's history of operations are recorded to construct a matrix, denoted as $\boldsymbol{A}$, to maintain the links between the user and his/her posts. Rows of $\boldsymbol{A}$ denote posts, and of $\boldsymbol{A}$ columns denote users. As shown in Fig. 2, the TD-HITS method creates a direct link between users and their posts, with regard to the corresponding individual user's operations.

In addition, in this paper, we extend the HITS algorithm to exploit the inseparable connection between users and their corresponding posts for the purpose of extracting only high-quality posts and influential users.

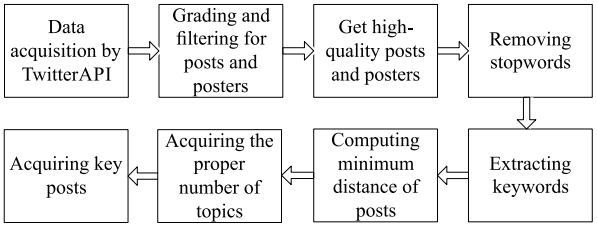


Fig. 1 TD-HITS procedure.

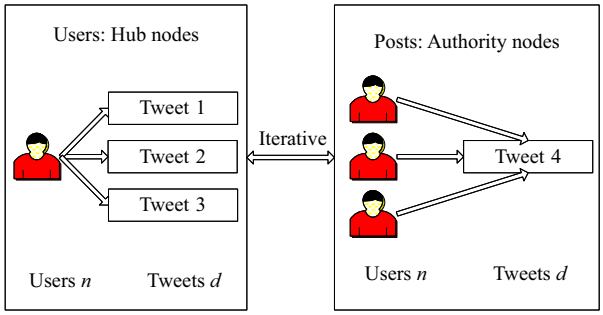


Fig. 2 Iterative model for determining the authority score of posts and the hub score of users.

Thus, the proposed TD-HITS method can effectively filter out random low-quality posts and ordinary users, and thereby avoid a phenomenon known as a bump, which generally reduces the efficiency and accuracy of event detection as well as the identification of influential spreaders.

Every post is given an authority score that indicates its significance. Similarly, each user is given a hub score denoting his/her influence. The most important feature of the iterative model is the mutual reinforcement of the relationship between the quality of a post and a user's real influence. For instance, a user who has published or forwarded many high-quality posts is more likely to make greater contributions than others in spreading a real-life event. Equally, a post that is forwarded by many highly influential users is more likely to be a high-quality post. A user's influence can be computed by calculating the sum of the authority scores (i.e., quality) of all posts published or commented upon by that user, and the quality of a post can be represented by the sum of the hub scores (or quality) of all the users who have forwarded that post. Then, the final scores can be iteratively calculated using Eqs. (1) and (2) for each post and user, respectively:

$$\text{n.h} = \sum \text{d.a} \tag{1}$$

$$\text{d.a} = \sum \text{n.h} \tag{2}$$

where d.a denotes post d 's authority score and n.h denotes user n's hub score[13]. The iterative processes for generating the final results are as follows:

$$A_n = \boldsymbol{M}^{\mathrm{T}} \cdot \boldsymbol{M} \cdot A_{n-1} \tag{3}$$

$$H_n = \boldsymbol{M} \cdot \boldsymbol{M}^{\mathrm{T}} \cdot N_{n-1} \tag{4}$$

where A_n and H_n denote the authority and hub scores at the n-th iteration, respectively, and $\boldsymbol{M}$ denotes the user–post matrix[16].

Thus, the TD-HITS method of filtration based on authority scores yields two significant advantages. First, the accuracy of event detection is improved significantly. Moreover, the time spent on event detection is significantly reduced. Consequently, the TD-HITS method is more efficient and effective than many existing methods, which usually ignore the processing of ordinary posts.

3.1.2 Extracting high-quality users based on the HITS algorithm

A high-quality post attracts the attention of many highly influential users and, typically, highly influential users post many high-quality posts. Intuitively, we can say that highly influential users publish or comment more high-quality posts than regular users. In addition, high-quality posts draw an increased level of attention from highly influential users, who spread or broadcast such posts in microblogging networks. The authority value of posts has been attributed more importance in the identification of influential spreaders, as has the hub value of users. Furthermore, special emphasis has been given to the theory that highly influential users are likely to publish many high-quality posts. Thus, filtration based on the hub score offers two significant advantages. First, the accuracy of the identification of influential spreaders is improved significantly. Second, the time cost is significantly reduced.

3.2 Topic decision method based on the authority and minimum distance of posts

Existing event detection models are weak with respect to determining the number of topics, which greatly reduces their efficiency and accuracy in event detection. To address this issue, a topic decision method[22] encompassing the authority value and the minimum distance between posts is essential for determining the importance of posts and the topic differences between posts. This can be achieved based on the idea that key posts related to an event usually have higher authority and there is often an evident topic difference between posts in microblogging networks. Thus, the number of topics can be automatically detected by the topic decision method. Finally, we use the number of topics as the "start" parameter of the LDA to detect hot events. Specifically, the TD-HITS method assumes that key posts have the following properties.

(1) Authority. Key posts in microblogging networks are surrounded by non-key posts. Similar to prototype-based clustering, where each cluster has a prototype, each topic in our proposed TD-HITS method is regulated by a topic-leading post. These topic-leading posts have a higher degree of influence regarding their respective topic, which also reflects their higher authority. As such, a post with high authority is more likely to be chosen as a topic.

(2) Topic dispersion. The topics of key posts differ in a microblogging network. As each topic is characterized by a key post, key posts are typically evenly distributed in a microblogging network. Thus, a key post will be separated by a larger topic-similarity distance from other posts with a different topic.

To capture the key posts characterized by the above two properties, a topic decision method is introduced[22], in which one dimension evaluates the "authority" properties of key posts and the other dimension characterizes the "topic dispersion" properties of key posts. Thus, K posts distributed in the right upper part of a graph can be automatically selected as key posts. Then, the LDA topic model can be used to cluster posts in a microblogging network into different topics. To execute the above process, the topic decision method must employ a minimum distance measurement to describe the dispersion between key posts.

In our proposed method, first, we compute the minimum distance between each key post and other posts of higher authority values. Then, we select the number of topics as the "start" parameter of the LDA. Among these carefully selected topics, posts in the microblogging network are clustered by LDA and Gibbs sampling.

The minimum distance δ_i $(i = 1, 2, \ldots, n)$ is calculated by computing the distance between post V_i and other posts with higher authority, as follows:

$$\xi_i = \min(d_{ij}), \quad j : A_j > A_i \tag{5}$$

where d_{ij} is the Euclidean distance between post V_i and post V_j.

To compute the Euclidean distance, we must first transform n posts of a post network into points in the same spatial area. One simple way to do so is to calculate the pairwise similarities of the posts. Various similarity methods[23], such as Jaccard similarity[24] and signal similarity[25], can be used to calculate the pairwise similarities. In this study, we calculated the similarity of posts based on the signal similarity, since we found signal similarity to perform better than Jaccard similarity, based on our experimental results.

Signal similarity is defined by the signaling propagation process. Each post is regarded as an initial signal source that excites the whole network one time. All of the other users record the number of signals they receive. During each step, posts send all of their signals to their neighbors as well as to themselves. After t steps, the signal distribution of a given post in comparison to that of other posts can be taken as the influence of this source post on the entire post network. Generally speaking, the source post influences its own topic network first and then affects other posts by spreading signals. Obviously, posts in the same topic network have similar effects on other posts. Then, the i-th column **INF** indicates the effect of post V_i on the entire network in t steps. Thus, we can obtain n vectors $\mathbf{INF}_1, \mathbf{INF}_2, \ldots, \mathbf{INF}_3$ in Euclidean space. This process can be expressed as follows:

$$\mathbf{INF} = (\boldsymbol{A} + \boldsymbol{I})^t \tag{6}$$

where $\boldsymbol{I}$ is an n-dimensional identity matrix and t is the total steps taken in the signaling propagation ($t = 3$ in implementation). Since post networks are usually sparse, the computation of **INF** is not time-consuming.

Using the above equations, we can calculate the similarity of all posts in the post network. It is easy to see that the authority value of each post is modified by the distance of the corresponding post from other posts in the post network. This suggests that posts located closer to other posts in the network obtain greater rank values, which indicates that these posts are more important than surrounding posts.

Specifically, if there are some posts with the same authority value, posts with a smaller post ID are ranked higher. For a post V_k with maximal rank value, $A_k = \max(A_i)$, obviously, V_k is more prominent than its neighbors, and has the greatest likelihood of being chosen as a key post. Hence, we assign its minimum distance δ_k as follows:

$$\xi_k = \max(\xi_i), \quad i \neq k \tag{7}$$

By this minimum distance, the values of different posts vary distinctly, and high-quality posts can be easily identified according to their rank value in a microblog network. Furthermore, based on the assumptions of our model, key posts are those having higher authority values and are dispersed throughout the microblogging network. In summary, we use this topic decision method in 2-dimensional space to automatically detect the number of topics, wherein one dimension is the authority value of the posts and the other is the minimum distance between the posts, as defined above. Therefore, in the proposed topic-decision method, posts located in the right upper co-ordinates are identified as key posts.

4 TS-LDA Model

The TS-LDA model has two modules. First, with the number of topics selected according to the TD-HITS method, posts in the post network are clustered according to the LDA topic model and Gibbs sampling. Second, the influential spreaders in events are identified by the hub value in the user–post network and local features in the user–user network. Figure 3 illustrates

the process of event detection and the identification of influential spreaders.

4.1 LDA topic model and Gibbs sampling

4.1.1 LDA topic model

PLSA[6] and LDA[7] are both widely used topic models for extracting hidden variables from a collection of posts. There are some similarities between these two models in their detection of events from among a large number of posts.

The PLSA model employs the following steps to obtain the "post–word" generation model:

(1) Select a post d_i in accordance with a probability $P(d_i)$;

(2) Determine the topic distribution of the post after selecting post d_i;

(3) Choose an implicit topic category z_k according to the probability $P(z_k|d_i)$ from the topic distribution;

(4) Determine the distribution of words in the topic after the selection of z_k;

(5) Select a word w_j according to the probability $P\left(w_j|z_k\right)$ from the distribution of words.

Below, we demonstrate the way in which a post is generated using the LDA model, as compared with the method used in the PLSA model:

(1) Choose a post d_i according to the prior probability $P(d_i)$;

(2) From the Dirichlet distribution α, select the topics distribution θ_i of the generated post d_i. In other words, the topic distribution θ_i is generated by the super parameter α in the Dirichlet distribution;

(3) Sample the topic z_{ij} of the j-th word in the post d_i from the topic's polynomial distribution θ_i;

(4) The Dirichlet distribution generates the distribution of the words d;

(5) Sample the distribution of words in the final set of words.

From the above two processes, we can see that LDA adds two prior probabilities in the Dirichlet distribution for the topic and word distributions, in addition to those in the PLSA.

In other words, first, the LDA model incorporates two prior probabilities into the Dirichlet distribution while determining a post. However, PLSA directly determines the topics distribution. In addition, there is also a certain probability being considered in LDA to generate the topic distribution. Second, in PLSA, the distribution probability of the words is determined by the topic distribution. However, the LSA word distribution is generated by the Dirichlet distribution of the super parameter beta. That is, in the PLSA model, the probability that post D generates topic Z, and the probability that topic Z generates the word w are two fixed values. However, in LDA, the topic distribution (the probability distribution of each topic in the post), as well as the word distribution (the probability distribution of each word in a topic), are randomly generated single values, which can also differ and have several probabilities. Similar to the beta distribution, binomial and Dirichlet distributions are actually distributions of the polynomial conjugate prior probability distribution. However, the dependability of the entire model is based on Dirichlet's prior distribution. Hence, LDA is more appropriate and accurate than PLSA in detecting events.

Table 1 lists and defines the symbols used in the LDA model (Fig. 4).

Table 1 Symbols used in the LDA model.

Symbol	Meaning
α	Super parameter of θ_m
β	Super parameter of ϕ_k
d	Post
w	Word
z	Topic
θ	Topic distribution
ϕ	Word distribution
Z	Number of topics
K	Number of words

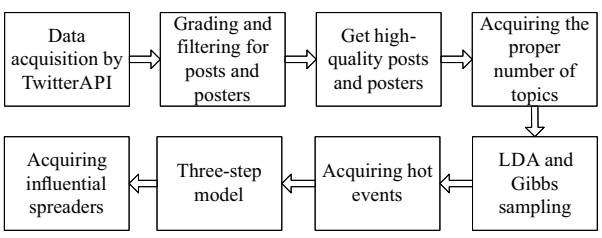


Fig. 3 Procedure of the TS-LDA model.

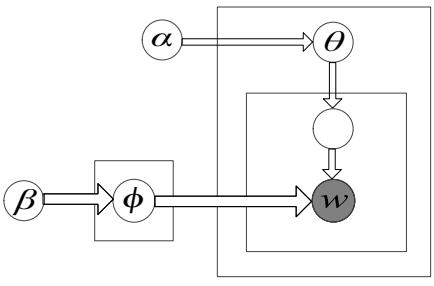


Fig. 4 LDA model[3].

$$P(d,w) = P(d)P(w|d) = P(d)\sum_z P(w|z)P(z|d) \tag{8}$$

where $P(d)$ represents the probability of selecting a post d from the post set, $P(z|d)$ is the probability of selecting a topic z from a post d, and $P(w|z)$ is the probability of selecting a word w from a topic z. As such, Eq. (2) can be expressed as follows:

$$P(d,w) = P(d)P(w|d) = \frac{1}{d}\sum_z \phi_{w|z}\theta_{z|d} \tag{9}$$

In the process of LDA modeling, both $P(d)$ and $P(w|d)$ can be easily computed, so the values of $\phi_{w|z}$ and $\theta_{z|d}$ can also be calculated. Finally, we can obtain the topic distribution in each post and the word distribution in each topic.

4.1.2 Gibbs sampling

The variables in the TS-LDA model cannot deliver an exact value that can be used for further estimation. Therefore, we adopt Gibbs sampling[26,27] to make an approximate inference of the variables. Gibbs sampling is a simple and widely applicable Markov-chain Monte Carlo algorithm. In comparison with the other inference methods used by latent variable models, such as variational inference and maximum likelihood estimation, Gibbs sampling exhibits two exceptionally advantageous features. First, it provides a reliable level of accuracy as it asymptotically approaches the correct distribution. Second, it is more memory-efficient since it need only maintain the counters and state variables, which makes it the preferred method for dealing with large-scale datasets. A more detailed comparison of these methods can be found in Ref. [26]. The basic concept underlying Gibbs sampling is the alternative estimation of parameters by replacing the value of one of the variables with a value drawn from the distribution of that variable, conditioned on the values of the remaining variables. In LDA, all three types of latent variables z, ϕ, and θ must be sampled. However, with the collapsed Gibbs sampling technique, ϕ and θ can be integrated due to the conjugate priors α and β. Consequently, we need only sample topic z. To perform Gibbs sampling, we first randomly choose the initial states of the Markov chain. Then, we calculate the conditional distribution $P(z|\alpha,\beta)$ for each z by applying the chain rule to the joint probability of the entire dataset. Thus, we can obtain the conditional probability conveniently, as follows:

$$P(z|\alpha,\beta) \propto \frac{(n_{w|z}+\beta)}{(\sum_w n_{w|z}+\beta)} \cdot \frac{n_{z|d}+\alpha}{\sum_z n_{z|d}+\alpha} \tag{10}$$

where $n_{w|z}$ is the number of words in the topic z and $n_{z|d}$ is the number of topics in the post.

We iteratively repeat Formula (10) and continue to sample all the topics until the sampling results are stabilized.

Finally, we can easily estimate the topic–word distribution ϕ and post–topic distribution θ as follows:

$$\theta_{z|d} = \frac{n_{z|d}+\alpha}{\sum_z n_{z|d}+\alpha} \tag{11}$$

$$\phi_{w|z} = \frac{n_{w|z}+\beta}{\sum_w n_{w|z}+\beta} \tag{12}$$

4.2 Identification of influential spreaders

In this section, we introduce our approach for identifying influential spreaders in the network. As noted above, high-quality posts can draw more attention from highly influential users, who spread or broadcast these posts in microblogging networks. Nevertheless, developing a method for identifying influential spreaders effectively and efficiently in microblogging networks has presented a considerable challenge. Many significant evaluation methods have been proposed to address this problem[28–30], including degree centrality, clustering coefficient centrality, and betweenness centrality[31,32].

However, degree centrality and the clustering coefficient centrality of spreaders can only characterize local network information. Moreover, computing betweenness centrality is highly complex due to the need to calculate the shortest path. Almost all of these methods employ only one centrality measure, and each method has its corresponding disadvantage and limitation.

The above centrality methods alone cannot be directly applied to tweet compositions between users and posts in user–post networks. Furthermore, there is a strong possibility that high-hub users publish a lot of high-quality posts. So, in user–post networks, there are too many users having the same hub value, which makes it impossible to rank users effectively.

To solve this problem, we propose a novel concept in which influential social network spreaders must satisfy one activity degree condition and one network topology condition within a certain time period: a high degree of activity and high number of local features. First, there is an expectation that users with a high degree of activity will spread information, ideas, or rumors very quickly

in the early stages of the spreading process. Second, the number of local features of users is measured by the sum of their neighbor connections. Here the expectation is that users with a high number of local features will trigger an early and rapid accumulation of contagious transmissions among a large number of candidate users. Finally, the degree of activity of users can be acquired from the hub values of users with posts. Also, the expectation here is that high-activity-degree users will spread information, ideas, or rumors much more quickly than regular users.

The three-step model shown in Fig. 5 illustrates our proposed method for obtaining the activity and local information of users in a microblogging network. In Step 1, the degree of activity is determined to analyze the global features of users in a microblogging network. The results are used to compute the final influence of these users. In Step 2, the degree centrality is used to measure local user features. In Step 3, the degree of activity and local features are combined to determine the influential spreaders.

Specifically, first, the hub value of users is used to obtain global information about users in a microblogging network. Second, a user's degree of centrality is used to analyze his/her local feature value in the microblogging network. A high influence value indicates that a user has high degree of centrality, which in turn indicates that the user is capable of reaching users in the widest possible local range. The local feature of user P is defined as follows:

$$\overline{P_i} = \frac{P_i - P_{\min}}{P_{\max} - P_{\min}} \tag{13}$$

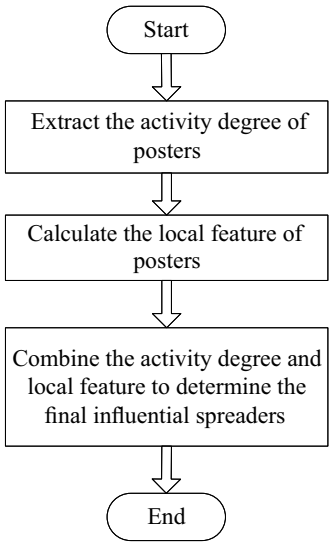


Fig. 5 Three-step model.

where P_i is the degree centrality of i and $\overline{P_i}$ is the normalized local feature required for the case under consideration.

Third, according to the definition of an influential spreader, the hub value and $\overline{P_i}$ are considered simultaneously to maximize the spreading capability of user i in the microblogging network. Finally, the degree of global activity H_i and local feature value $\overline{P_i}$ are combined to denote IF_i, the influence of node i, which is defined as follows:

$$\mathrm{IF}_i = H_i \times \overline{P_i} \tag{14}$$

Therefore, we address the problems identified above by combining the hub value of users with the degree of locality of users. In general, a user with more connections with neighbors is more likely to be an influential spreader in a user–user network. Inspired by this idea, we propose a novel influence measure that considers the hub and locality degree values of users. To a large extent, the method proposed in this paper improves accuracy in the identification of influential spreaders.

5 Experiments

In this section, we detail the experiments we conducted on real-world short-text collections to demonstrate the effectiveness of our proposed TD-HITS method and TS-LDA model. We consider two typical topic models as benchmark methods, namely PLSA and LDA.

In the rest of this section, we describe our collection of the dataset, experimental setup and analysis, the baseline approaches, and model evaluation.

5.1 Dataset

We generated our dataset from Twitter (http://twitter.com/) via Twitter API. This dataset consists of 40 000 posts from October 25–28, 2015. As discussed above, to reduce the impact of the bump phenomenon, we included only those users who published or commented upon posts in our dataset. After filtering unwanted users and posts, the dataset comprised 2139 high-quality posts and 1887 users.

5.2 Experimental settings

We conducted the experiments on a computer with an Intel I3 3.4 GHz CPU and 4 GB memory.

We tuned the parameters via a grid search. For PLSA, we fixed the mixture weight of the background model λ_B to 0.05[11]. For LDA, α = 0.5 and β = 0.1. In all the experiments, we used Gibbs sampling

for 1000 iterations and the Expectation–Maximization (EM) algorithm for 1000 iterations. The results reported here are the average of five runs. In the process of filtering high-quality posts, we set all of the initial authority scores d.a and hub scores n.h to 1.

5.3 Baseline approaches

We validated the improved efficiency and effectiveness of the proposed TS-LDA by evaluating our model against PLSA[6], LDA[7], and Efficient eVent dEtection (EVE)[33], which are classic latent semantic analysis algorithms.

5.4 Evaluation methods

(1) Number of topics: As shown in Fig. 6, from the experimental results, the top eight minimum distances of these posts are far greater than the others. Therefore, we chose eight as the LDA "start" parameter. With this carefully selected number of topics, we then clustered the posts by LDA and Gibbs sampling[26] to detect hot events efficiently and accurately, which address the shortcomings in many existing methods[8–11, 14–18, 21].

(2) Key posts of hot events: As shown in Table 2, when the authority values of posts are equal, they can be sorted according to the minimum distance of posts, which improves the accuracy of event detection. The TD-HITS method can detect the top eight key posts of hot events according to their minimum distances, in contrast to many existing methods[8–11, 14–18, 21].

(3) Event-detection effectiveness: As shown in Table 3, key posts are sorted according to their minimum distances. Meanwhile, the popular degree of different events in each time period and the development process of an event can be distinguished clearly by the creation time of the top eight key posts.

(4) Trend of reply number changes over time: As we can see in Fig. 7, the reply number of events for the top four popular events changes from the beginning to the peak and then to extinction. Existing event detection methods, such as the PLSA and LDA models,

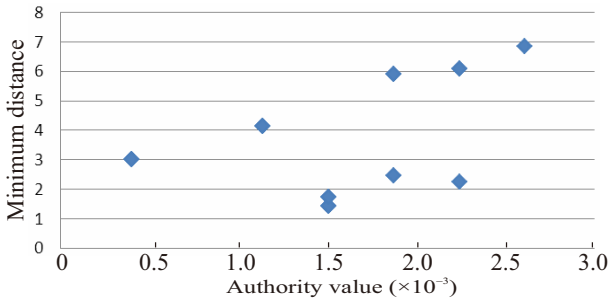


Fig. 6 Number of topics from the TD-HITS method.

Table 2 Minimum distance and authority of posts.

Key post ID	Minimum distance	Authority value
659051011055611904	6.855 654 600	2.609 579 007 761 16
659094561617133568	6.855 654 600	2.609 579 007 761 16
658903037663055872	6.082 762 530	2.241 254 882 530 98
658786650407964672	6.082 762 53	2.241 254 882 530 98
658750778572709889	5.916 079 783	2.241 254 882 530 98
658676220230545409	5.916 079 783	2.241 254 882 530 98
658675883633414145	5.916 079 783	2.241 254 882 530 98
658602777183154177	5.916 079 783	2.241 254 882 530 98

Table 3 Evaluation results for event detection.

Minimum distance	Real-life event	Key post	Created time
6.855 654 6	The rise and controversy of classical economics	@namasteacup "classical economics"is specifically in the text:p	Tue Oct 27 16:56:20 2015
6.855 654 6	Economic deficit in United States	@Shamsher1111 @johnefrancis If you have a master in economics and don't understand uses of deficit spending, you are a very great fool.	Tue Oct 27 19:49:23 2015
6.082 762 53	Economic crisis in Poland	@BeingAnkit_ My mind starts boggling at Economics. I better leave you to study.	Tue Oct 27 07:08:20 2015
6.082 762 53	The rise of cultural economics	"each part has a size measuring its efficiency economics became more efficient than culture for organizing society @enleuk"	Mon Oct 26 23:25:51 2015
5.916 079 783	The rise of cultural economics	@NYSELaxative @StartlinglyOkay What kind of input, output and filter? And economics is limited to property and only one part of culture.	Mon Oct 26 21:03:19 2015
5.916 079 783	The rise of football economics	@ArsenalReport @JanuzajA11 @Firzaapras err I study economics so I I'd know about this subject especially, and its a fact that it doesn't	Mon Oct 26 16:07:03 2015
5.916 079 783	The rise of football economics	@mk_9873 @januzaja11 @firzaapras Maybe because you only hang out with mouth breathers? It's how all economics work, not just football.	Mon Oct 26 16:05:42 2015
5.916 079 783	Economic crisis in Ireland	@UB_Economics I am sorry to hear this @hazeyhall, have you managed to arrange an appointment now? PH	Mon Oct 26 11:15:12 2015

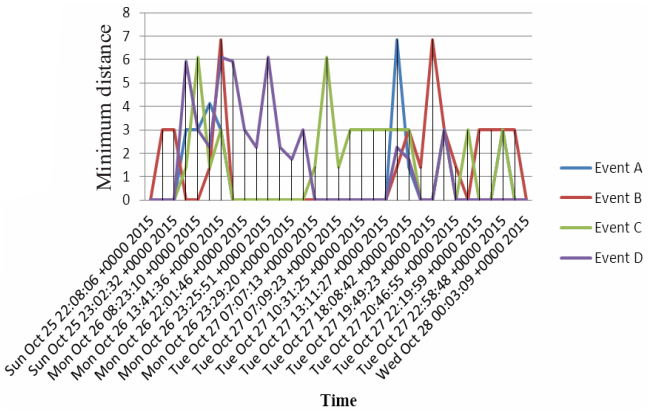


Fig. 7 Trends in reply number changes of top four popular events over time.

can determine the trend of the reply number of event changes over time, but the TS-LDA can determine this trend via the minimum distance of key posts about the same event, which creates a foundation for future research and the tracking of public events.

Since the TD-HITS method is based on the minimum distance and authority of posts, we can also compare their changes in the top four popular events over time from Fig. 7. We can also easily identify key posts of the hottest event, which can then be used to trace hot events, as shown in Fig. 8. The PLSA and LDA models usually cluster on the basis of tweets alone, so experience increased complexities in detecting the trends of event reply number changes over time. Also, they cannot distinguish between similar events.

(5) Event-detection precision and efficiency: To compare the precision and efficiency of our model with those of the PLSA, LDA, and EVE models, we evaluated effectiveness in our experiments. For PLSA, LDA, EVE, and our TS-LDA, we defined precision as follows:

$$\text{Precision} = \frac{a}{b} \tag{15}$$

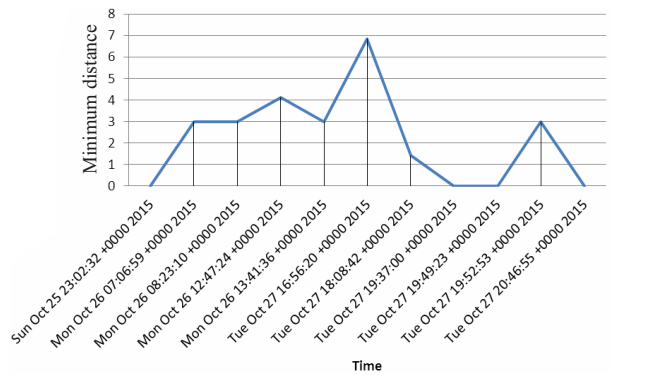


Fig. 8 Trend of reply number changes for the hottest event over time.

where a represents the number of detected events matching real life events, b is the total number of the events detected by the same algorithm.

Table 4 shows a comparison of the precision of the three methods, and Table 5 shows a comparison of their time efficiencies. As shown in these tables, our model can find seven events if K is set for 8. However, for the PLSA, LDA, and EVE models, if K is set any greater than 6, then these models can detect all the events only by artificial selection. If K is set to 1 or 5, then all events would remain undetected. At the same time, if the K value is greater than 8, such as 10, then although these models can detect all of the events, their time efficiencies are very low. Therefore, our proposed TS-LDA model is both accurate and efficient, and its effectiveness is better than those of the PLSA, LDA, or EVE models.

(6) Results of influential spreader identification: As noted above, influential spreaders are identified based on the hub value in the user–post network, and the degree value in the user–user network. Thus, we consider both post and user information for hot events while also identifying influential spreaders for related events. Also, we know from the three-step model that if some users have the same hub value, users with a higher degree value are ranked higher. Table 6 shows the top five results of the TS-LDA model for the identification of influential spreaders. We can also see that user 80864710 is the screen name of a user having the same hub value, but a lower degree value than user 29442313. Hence, user 29442313 ranks higher according to the three-step model. In addition, Table 7 shows the top

Table 4 Comparison of precision.

Method	$K = 1$	$K = 5$	$K = 8$	$K = 10$
PLSA	1/1	5/5		6/10
LDA	1/1	5/5		6/10
EVE	1/1	5/5		6/10
TS-LDA			6/8	

Note: K is the number of possible events.

Table 5 Comparison of time efficiency.

Method	Time ($K = 8$)				
	HITS	Topic decision method	Gibbs sampling	EM	Total
PLSA	N.A	N.A	N.A	24.05 min	24.05 min
LDA	N.A	N.A	15.62 min	N.A	15.62 min
EVE	10922 ms	N.A	N.A	7.32 min	7.51 min
TS-LDA	10922 ms	3.69 min	1.16 min	N.A	5.03 min

Note: K is the number of possible events.

Table 6 Results of identification of influential spreaders.

User ID	Hub value	Degree value
1410108115	0.003402463	0.0015
29442313	0.003078419	0.0019
80864710	0.003078419	0.0012
25073877	0.002754375	0.0125
25654421	0.002106286	0.0017

Table 7 Results of identification of influential spreaders in related event.

User ID	Related event
1410108115	The rise of cultural economics
29442313	Economics deficit in United States
80864710	Economics crisis in Poland
25073877	Economic deficit in United States
25654421	The rise of football economics

five results of the TS-LDA model for the identification of influential spreaders in a related event, which can facilitate the tracking of related events in a timely and efficient manner.

(7) Effectiveness of the identification of influential spreaders: We can confirm the effectiveness of the identification of influential spreaders by counting retweets and comments, which represent the breadth and depth of influence. Hence, according to Table 8, the second and third users have the same retweet and comment counts, which indicate that they have the same influence in this period of time. Also, the top five influential spreaders identified by the TS-LDA model all have higher retweet and comment counts than the other users. Hence, our proposed TS-LDA model is effective, accurate, and dependable. Moreover, the influential spreaders identified in the related event by our TS-LDA model are important factors in the dissemination of hot events.

6 Conclusion and Future Work

In this paper, we proposed the HITS-based topic-decision method, TD-HITS. This proposed approach creates a smaller, high-quality training data set by filtering high-quality posts and high-quality users from a collection of users and posts. This approach largely reduces the impact of unrelated posts and occasional users, thereby improving the efficiency and accuracy of the event detection process. Moreover, this approach can automatically detect the correct number of topics and identifies event-related key posts to realize higher precision. In addition, we also proposed an LDA-based three-step model TS-LDA, which detects critical events by analyzing the number of topics and identifying the influential spreaders linked to them. This approach utilizes both post and user information, which can enable a better understanding in a timely and accurate manner of the users involved in these critical incidents.

Table 8 Effectiveness of identification of influential spreaders in related event.

User ID	Retweet and comment count
1410108115	20
29442313	18
80864710	18
25073877	15
25654421	12

Our experimental results for a Twitter dataset demonstrate the effectiveness of our proposed methods in event detection, key post detection, and the identification of influential spreaders. In particular, it excels in detecting the trend in the number of event changes over time. In future work, to better understand the transmission and control of events, we plan to further investigate the behaviors of influential spreaders and develop a dynamic community detection model that can evolve over time.

Acknowledgment

The work was supported by the National Natural Science Foundation of China (Nos. 61502209 and 61502207), the Natural Science Foundation of Jiangsu Province of China (No. BK20130528) and Visiting Research Fellow Program of Tongji University (No. 8105142504).

References

[1] X. M. Zhou and L. Chen, Event detection over twitter social media streams, *VLDB J.*, vol. 23, no. 3, pp. 381–400, 2014.

[2] A. Aldhaheri and J. Lee, Event detection on large social media using temporal analysis, in *Proc. 7th Annu. Computing and Communication Workshop and Conf.*, Las Vegas, NV, USA, 2017, pp. 1–6.

[3] P. Yan, MapReduce and semantics enabled event detection using social media, *J. Artif. Intell. Soft Comput. Res.*, vol. 7, no. 3, pp. 201–213, 2017.

[4] Y. D. Zhou, H. Xu, and L. Lei, Event detection based on interactive communication streams in social network, in *Proc. 9th EAI Int. Conf. Mobile Multimedia Communications*, Xi'an, China, 2016, pp. 54–57.

[5] T. Hofmann, Probabilistic latent semantic indexing, in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Research and*

Development in Information Retrieval, Berkeley, CA, USA, 1999, pp. 50–57.

[6] T. Hofmann, Probabilistic latent semantic indexing, in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, Berkeley, CA, USA, 1999, pp. 50–57.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet allocation, *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[8] Q. M. Diao, J. Jiang, F. D. Zhu, and E. P. Lim, Finding bursty topics from microblogs, in *Proc. 50th Annu. Meeting of the Association for Computational Linguistics: Long Papers–Volume 1*, Jeju Island, Korea, 2012, pp. 536–544.

[9] X. H. Wang, C. X. Zhai, X. Hu, and R. Sproat, Mining correlated bursty topic patterns from coordinated text streams, in *Proc. 13th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Jose, CA, USA, 2007, pp. 784–793.

[10] L. AlSumait, D. Barbara, and C. Domeniconi, On-Line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking, in *Proc. 8th IEEE Int. Conf. Data Mining*, Pisa, Italy, 2008, pp. 3–12.

[11] J. X. Li, Z. Y. Tai, R. C. Zhang, W. R. Yu, and L. Liu, Online bursty event detection from microblog, in *Proc. 7th IEEE/ACM Int. Conf. Utility and Cloud Computing*, London, UK, 2014, pp. 865–870.

[12] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, Automatic resource compilation by analyzing hyperlink structure and associated text, *Comput. Netw. ISDN Syst.*, vol. 30, nos. 1–7, pp. 65–74, 1998.

[13] J. Bao, Y. Zheng, and M. F. Mokbel, Location-based and preference-aware recommendation using sparse geo-social networking data, in *Proc. 20th Int. Conf. Advances in Geographic Information Systems*, Redondo Beach, CA, USA, 2012, pp. 199–208.

[14] J. Kleinberg, Bursty and hierarchical structure in streams, in *Proc. 8th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (*KDD '02*), Edmonton, Canada, 2002, pp. 91–101.

[15] Y. M. Yang, T. Pierce, and J. Carbonell, A study of retrospective and on–line event detection, in *Proc. 21st Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval* (*SIGIR '98*), Melbourne, Australia, 1998, pp. 28–36.

[16] M. Mathioudakis and N. Koudas, Twittermonitor: Trend detection over the twitter stream, in *Proc. 2010 ACM SIGMOD Int. Conf. Management of Data*, Indianapolis, IN, USA, 2010, pp. 1155–1158.

[17] J. Allan, V. Lavrenko, D. Malin, and R. Swan, Detections, bounds, and timelines: UMass and TDT–3, in *Proc. Topic Detection and Tracking Workshop*, *TDT*–3, Vienna, Austria, 2000, pp. 167–174.

[18] F. Atefeh and W. Khreich, A survey of techniques for event detection in twitter, *Comput. Intell.*, vol. 31, no, 1, pp. 132–164, 2015.

[19] Twitter, REST API v1.1 resources, https://dev.twitter.com/docs/api/1.1/, 2017.

[20] Facebook, Quickstart for the Azure AD Graph API, https://docs.microsoft.com/en–us/azure/active–directory/develop/active–directory–graph–api–quickstart, 2017.

[21] J. S. Weng and B. S. Lee, Event detection in *Twitter*, in *Proc. 5th Int. AAAI Conf. Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 401–408.

[22] Y. F. Li, C. Y. Jia, and J. Yu, A parameter–free community detection method based on centrality and dispersion of nodes in complex networks, *Phys. A: Stat. Mech. Appl.*, vol. 438, pp. 321–334, 2015.

[23] L. Y. Lü and T. Zhou, Link prediction in complex networks: A survey, *Phys. A: Stat. Mech. Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011.

[24] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et du Jura, *Bulletin del la Société Vaudoise des Sciences Naturelles*, vol. 37, no. 142, pp. 547–579, 1901.

[25] Y. Q. Hu, M. H. Li, P. Zhang, Y. Fan, and Z. R. Di, Community detection by signaling on complex networks, *Phys. Rev. E*, vol. 78, no. 1, p. 016115, 2008.

[26] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, On smoothing and inference for topic models, in *Proc. 25th Conf. Uncertainty in Artificial Intelligence*, Montreal, Canada, 2009, pp. 27–34.

[27] R. Alhamzawi and K. M. Yu, Variable selection in quantile regression via Gibbs sampling, *J. Appl. Stat.*, vol. 39, no. 4, pp. 799–813, 2012.

[28] P. G. Sun and Y. Yang, Methods to find community based on edge centrality, *Phys. A Stat. Mech. Appl.*, vol. 392, no. 9, pp. 1977–1988, 2013.

[29] M. G. Campiteli, A. J. Holanda, L. D. H. Soares, P. R. C. Soles, and O. Kinouchi, Lobby index as a network centrality measure, *Phys. A: Stat. Mech. Appl.*, vol. 392, no. 21, pp. 5511–5515, 2013.

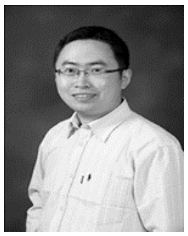
[30] J. Sohn, D. Kang, H. Park, B. G. Joo, and I. J. Chung, An improved social network analysis method for social networks, in *Advanced Technologies, Embedded and Multimedia for Human–Centric Computing*, Y. M. Huang, H. C. Chao, D. J. Deng, and J. J. Park, eds. Amsterdam, The Netherlands: Springer, 2014, pp. 115–123.

[31] P. Bonacich, Factoring and weighting approaches to status scores and clique identification, *J. Math. Sociol.*, vol. 2, no. 1, pp. 113–120, 1972.

[32] O. Green and D. A. Bader, Faster betweenness centrality based on data structure experimentation, *Procedia Comput. Sci.*, vol. 18, pp. 399–408, 2013.

Leilei Shi received the BS degree from Nantong University, China, in 2012, and the MS degree from Jiangsu University, China, in 2015. He is currently working towards the PhD degree at the School of Computer Science and Telecommunication Engineering, Jiangsu University, China. His research interests include event detection, data mining, social computing, and cloud computing.

Yan Wu received the MS degree from Shandong University of Science and Technology, China, in 2009, and the PhD degree from Tongji University, China, in 2014. He is currently a lecturer with the School of Computer Science and Telecommunication Engineering in Jiangsu University, China. His research interests include formal methods, service-oriented computing, and cloud computing.



Lu Liu is the professor of distributed computing in the University of Derby, UK and adjunct professor in Jiangsu University, China. Prof. Liu received the PhD degree from University of Surrey in 2007 and MS degree from Brunel University, in 2003. Prof. Liu's research interests are in areas of cloud computing, social computing, service oriented computing, and peer-to-peer computing. He is the Fellow of British Computer Society (BCS).

Xiang Sun received the BS degree from Jiangsu University, China, in 2013, and the MS degree from Jiangsu University, China, in 2016. He is currently working towards the PhD degree in the University of Derby, UK. His research interests include event detection, data mining, and social computing.

Liang Jiang received the BS degree from Nanjing University of Posts and Telecommunications, China, in 2007, and the MS degree from Jiangsu University, China, in 2011. He is currently working towards the PhD degree at the School of Computer Science and Telecommunication Engineering, Jiangsu University, China. His research interests include OSNs, computer networks, and network security.