9-2016

Detecting community pacemakers of burst topic in Twitter

Guozhong DONG
Harbin Engineering University

Wu YANG
Harbin Engineering University

Feida ZHU
Singapore Management University, fdzhu@smu.edu.sg

Wei WANG
Harbin Engineering University

Published in APWeb 2016: Proceedings of the 18th Asia Pacific Web Conference: Suzhou, China, September 23-25. pp. 245-255. https://doi.org/10.1007/978-3-319-45814-4_20

Detecting Community Pacemakers of Burst Topic in Twitter

Guozhong Dong[1], Wu Yang[1(✉)], Feida Zhu[2], and Wei Wang[1]

[1] Information Security Research Center,
Harbin Engineering University, Harbin, China
yangwu@hrbeu.edu.cn
[2] Singapore Management University, Singapore, Singapore

Abstract. Twitter has become one of largest social networks for users to broadcast burst topics. Influential users usually have a large number of followers and play an important role in the diffusion of burst topic. There have been many studies on how to detect influential users. However, traditional influential users detection approaches have largely ignored influential users in user community. In this paper, we investigate the problem of detecting community pacemakers. Community pacemakers are defined as the influential users that promote early diffusion in the user community of burst topic. To solve this problem, we present DCPBT, a framework that can detect community pacemakers in burst topics. In DCPBT, a burst topic user graph model is proposed, which can represent the topology structure of burst topic propagation across a large number of Twitter users. Based on the model, a user community detection algorithm based on random walk is applied to discover user community. For large-scale user community, we propose a ranking method to detect community pacemakers in each large-scale user community. To test our framework, we conduct the framework over Twitter burst topic detection system. Experimental results show that our method is more effective to detect the users that influence other users and promote early diffusion in the early stages of burst topic.

Keywords: Twitter · Burst topic · User graph model · Community pacemakers

1 Introduction

With the development of social media, Twitter has been an important medium for providing the rapid spread of burst topic. When breaking news or events occur, influential users can post tweets about breaking news and share with their friends. Due to large number of people that have different user interests participating in conversation and discussion, some tweets spread among Twitter users and become the source of burst topics. As such, the main cause of burst topic is the information diffusion in user community. Figure 1 illustrates the

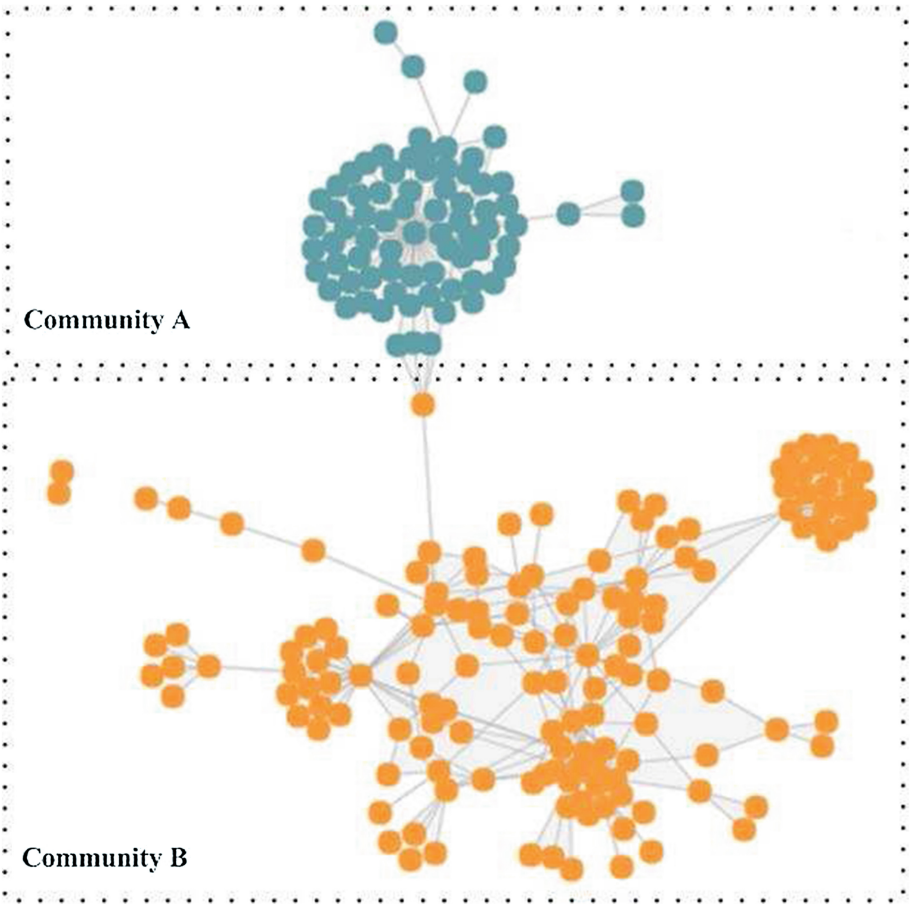

Fig. 1. Example of user community of burst topic

user community of burst topics detected by CLEar system[1], in which two main user communities are marked in different colors. So far, plenty of works focus on the influential users who are popular or famous in burst topics. However, these famous influential users are not the early spreader of burst topic that also influence their followers to spread the topic. In quite a lot of scenario, it is more important to detect the cause of burst topic diffusion in different user communities, which are called community pacemakers in this paper.

Unfortunately, detecting community pacemakers in burst topic has not been solved by the existing works. In this paper, we propose DCPBT (Detecting Community Pacemakers in Burst Topics) framework and implement the framework on CLEar system. When new burst topics are detected by CLEar system, DCPBT applies burst topic user graph construct algorithm to conduct user graph for each

[1] http://research.pinnacle.smu.edu.sg/clear/.

burst topic. Based on burst topic user graph, a user community detection algorithm based on random walker is proposed to detect user community in burst topic, which can adjust the number of user community adaptively and select large-scale user community. For large-scale user community, we propose a ranking method to detect community pacemakers in each large-scale user community. To summarize, the contributions of our work are listed as follows:

(1) We propose a burst topic user graph model which can represent the topology structure of burst topic propagation across a large number of Twitter users. In the burst topic user graph, nodes represent the burst topic users and edges represent the follower/followee relationship between users.
(2) A community pacemakers detection algorithm is proposed to detect community pacemakers in each large-scale user community of burst topic, which is more effective to detect the users that influence other users and promote early diffusion in the early stages of burst topic.
(3) We implement DCPBT framework on CLEar system, which can demonstrate the effectiveness of DCPBT framework.

The rest of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents the framework of DCPBT. Section 4 describes the experimental results. Finally, we conclude our work in Sect. 5.

2 Related Work

The study of burst topic [1–8] and user influence [9–17] have been studied in the last decade. As there are numerous research works focusing on it, here we introduce the ones most related to our work.

Burst Topic Detection: Prasadet et al. [1] propose a framework to detect emerging topics through the use of dictionary learning. They determine novel documents in the stream and subsequently identify topics among the novel documents. Agarwal et al. [2] model emerging events detection problem as discovering dense clusters in highly dynamic graphs and exploit short-cycle graph property to find dense clusters efficiently in microblog streams. Alvanaki et al. [3] present the "en Blogue" system for emergent topic detection. En Blogue keeps track of sudden changes in tag correlations and presents tag pairs as emergent topics. Takahashi et al. [4] apply a recently proposed change-point detection technique based on Sequentially Discounting Normalized Maximum Likelihood (SDNML) coding to detect abnormal messages and detect the emergence of a new topic from the anomaly measured through the model. Wang, Liu et al. [5] propose a system called SEA to detect events and conduct panoramic analysis on Weibo events from various aspects. Xie et al. [6,7] present a real-time system to provide burst event detection, popularity prediction, and event summarization. Shen et al. [8] analyze different burst patterns and propose real-time burst topics detection oriented Chinese microblog stream. The method detect burst entities

and cluster them to burst topics without requiring Chinese segmentation, which can obtain related messages and users at the same time.

User Influence: Cha et al. [9] analyze the influence of Twitter users by employing three measures that capture different perspectives: indegree, retweets, and mentions. They find that influence is not determined by single factor, but through many factors. Lee et al. [10] propose a method to find influentials by considering both the link structure and the temporal order of information adoption in Twitter. Weng et al. [11] propose an extension of PageRank algorithm to measure the influence of users in Twitter, which measures the influence taking both the topical similarity between users and the link structure. Brown et al. [12] investigate a modified k-shell decomposition algorithm based on user relationship to compute user influence on Twitter. Fang et al. [13] develop a novel Topic-Sensitive Influencer Mining (TSIM) framework in interest-based social media networks to find topical influential users and images. Saez-Trumper et al. [14] propose a ranking algorithm to detect trendsetters in information networks. The algorithm can identify persons that spark the process of disseminating ideas that become popular in the network.

Note that previous studies mainly aim at detecting burst topics and influential users. Different from other works, we consider the role of user community in burst topic diffusion and propose the problem of detecting community pacemakers in burst topics. We focus on detecting influential users that promote early diffusion in the user community of burst topic.

3 Framework of DCPBT

The framework of DCPBT that construct on CLEar system is shown in Fig. 2, which contains three functional layers, namely Data Layer, Model Layer and Presentation Layer.

The Data Layer provides two databases for efficient data storage and data query. The first one is to store burst topics detected by CLEar system, and provide query operation for new burst topic monitor module (NBTM) in Model Layer. The second one is to store Twitter stream data, which stores necessary data involved in burst topics. The Model Layer utilizes several important modules to detect community pacemakers in burst topics. NBTM monitors new burst topics via polling burst topic database. Once new burst topics are detected, NBTM sends burst topic data collect command to burst topic data collect module (BTDC). BTDC retrieves burst topic data from Hadoop cluster, constructs burst topic user graph for further processing. In order to detect pacemakers in burst topic, community pacemakers detection algorithm based on burst topic user graph is proposed. The Presentation Layer presents the user engagement series and pacemakers detected by DCPBT with a user-friendly interface.

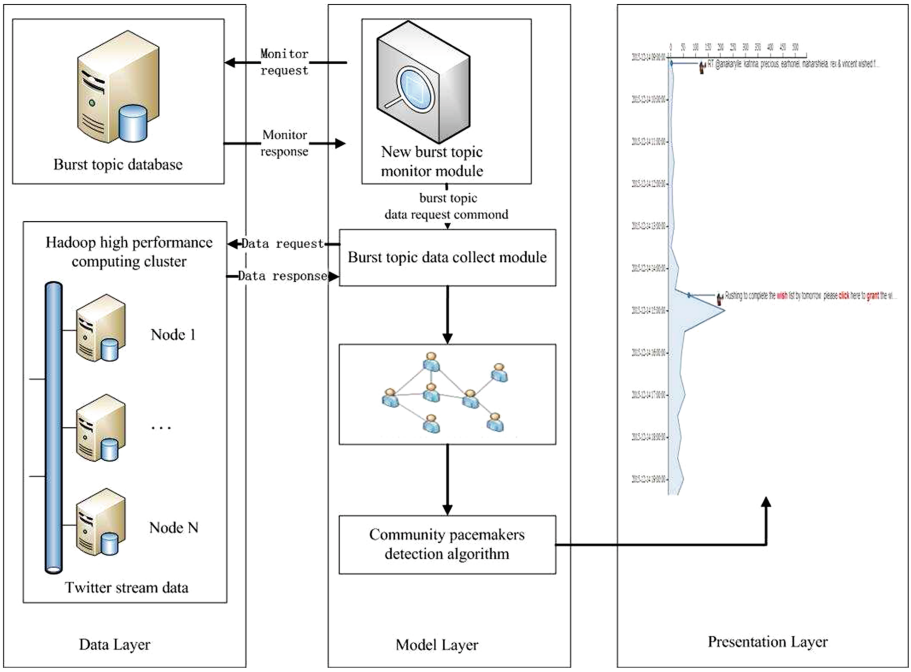

Fig. 2. The framework of DCPBT

4 Methods

In this section, we introduce some important models employed by DCPBT.

4.1 Burst Topic User Graph Model

Once a user posts a tweet related to the burst topic in Twitter, the tweet can spread to the user's followers, then followers who are interested in the burst topic may post or retweet the message. In order to represent the topology structure of burst topic propagation across a large number of Twitter users, in the burst topic user graph model, nodes represent the burst topic users and edges represent the follower/followee relationship between users.

The burst topic user graph of burst topic k can be formally defined as $G_k =< V_k, E_k, T_k >$. In detail, $V_k = \{u, \cdots, v, \cdots\}$ is the set of Twitter users over burst topic k, E_k represents the set of edges among Twitter users, in which a directed edge (u, v) means that u is the follower of v. $T_k = \{t(u), \cdots, t(v), \cdots\}$ is the earliest post time set of users over burst topic k.

By considering time information, the directed edges in topic user graph model can represent the direction of information flow and play a key role in detecting pacemakers, so we include time information in the edge weight. For each

$(u, v) \in E_k$, the edge weight $w(u, v)$ and the normalization of edge weight $W(u, v)$ can be defined as follows

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

$$W(u,v) = \frac{w(u,v)}{\sum\limits_{m \in Out(u)} w(u,m)} \tag{2}$$

where $Out(u)$ is the following set of user u in burst topic k.

For each burst topic, BTDC collects burst topic dataset from CLEar system and constructs burst topic user graph model. Given a tweet list TL of burst topic k, we first sort the tweet in descending order by post time and in it burst topic user set V_k. Afterwards, for each tweet in TL, we update burst topic user set. Finally, for each topic user in V_k, burst topic user graph are generated based on followee relationship and topic time of topic user.

4.2 Community Pacemakers Detection Algorithm

In this section, we introduce the community pacemakers detection algorithm in DCPBT. Based on the burst topic user graph, we use a random walker as a proxy, which can discover user community through minimizing a map equation over burst topic user graph. The map equation is first introduced in ref. [18]. Given a burst topic user graph with n users, the conditional probability that the random walker steps from user u to user v is given by the edge weight:

$$p_{u \to v} = W(u,v) \Big/ \textstyle\sum_v W(u,v) \tag{3}$$

To ensure the independent of the random walker starts in burst topic user graph, we use smart teleportation scheme and only record steps along links [19]. The stationary distribution is given by p_u^*, which can be expressed:

$$p_u^* = (1-\tau) \sum_v p_v^* p_{v \to u} + \tau \frac{\sum_v W(u,v)}{\sum_{u,v} W(v,u)} \tag{4}$$

The unrecorded visit rates on edge $q_{v \to u}$ and nodes p_u can now be formalized as follows:

$$q_{v \to u} = p_v^* p_{v \to u} \tag{5}$$

$$p_u = \sum_v q_{v \to u} \tag{6}$$

We use C to denote the community partition of burst topic user graph into m modules, with each node u assigned to a community i. m community codebooks and one index codebook are used to describe the random walker's movements within and between communities. The community transition rates $q_{i\leftarrow}$ and $q_{i\rightarrow}$

represent that the random walker enter and exit community i, which can be expressed by unrecorded visit rates on edge:

$$q_{i\leftarrow} = \sum_{u\in j\neq i, v\in i} q_{u\to v} \tag{7}$$

$$q_{i\to} = \sum_{u\in i, v\in j\neq i} q_{u\to v} \tag{8}$$

The map equation that can measure the per-step theoretical lower limit of a modular description of a random walker on user graph is given by:

$$L(M) = q_{\leftarrow} H(Q) + \sum_{i=1}^{m} R_i H(P^i) \tag{9}$$

Below we explain the terms of the map equation in detail. $L(M)$ represents the per-step description length for community partition, $q_{\leftarrow}$ represents the total probability that the random walker enters any of the m communities, which can be expressed:

$$q_{\leftarrow} = \sum\nolimits_{i=1}^{m} q_{i\leftarrow} \tag{10}$$

$H(Q)$ represents the frequency-weight average length of codewords in the index codebook, which is given by:

$$H(Q) = -\sum\nolimits_{i=1}^{m} (q_{i\leftarrow}/q_{\leftarrow}) \log(q_{i\leftarrow}/q_{\leftarrow}) \tag{11}$$

R_i represents the rate at which the community codebook i is used, which is given by:

$$R_i = \sum\nolimits_{u\in i} p_u + q_{i\to} \tag{12}$$

$H(P^i)$ represents the frequency-weight average length of codewords in community codebook i, which is given by:

$$H(P^i) = -(q_{i\to}/R_i) \log(q_{i\to}/R_i) - \sum\nolimits_{u\in i} (p_u/R_i) \log(p_u/R_i) \tag{13}$$

With the map equation, the burst topic user graph can be divided into different user communities. First, each user node is assigned to its own community. Then, in random order, each user node is moved to the neighboring community that results in the largest decrease of the map equation. If no move results in a decrease of the map equation, the user node stays in its original community. This procedure is repeated, each time in a new random order, until no move generates a decrease of the map equation. Then the network is rebuilt, with the communities of the last level forming the nodes at this level, and, exactly as at the previous level, the nodes are joined into communities. This hierarchical rebuilding of the network is repeated until the map equation cannot be reduced further.

At last, each community C_i represents user community in burst topic. The large scale user community set, denoted by C_l, is selected by:

$$C_l = \{C_i \,||C_i| \geq n/5\} \tag{14}$$

The community detection algorithm can adjust the number of user community adaptively, in which community number parameter is not needed. For large-scale user community, we propose a ranking method to detect community pacemakers in each large-scale user community. The pacemaker weight of user v in each large scale user community $C_l \in C$, denoted by $PM_{C_l}(v)$, is given by

$$PM_{C_l}(v) = dD(v) + (1-d) \sum_{u \in IN_{C_l}(v)} PM_{C_l}(u)W(u,v),\ 0 \leq d \leq 1 \tag{15}$$

where d is the damping factor, $IN_{C_l}(v)$ is the follower set of user v in user community C_l and $D(v)$ is a probability distribution over C_l. The distribution is topic dependent and is set to $1/|C_l|$ for all $v \in C_l$. The community pacemakers of user community C_l are the top N_l pacemaker weight of users in user community C_l.

5 Experiments

In order to test the advantage of community pacemakers detection algorithm, we have implemented and conducted a set of experiments on CLEar system. In this section, we first describe the dataset used in the experiments, and then present the experimental evaluation. The goal of experiment is to prove that our framework is more efficient than other approaches. In each experiment, we compare our PM ranking with TS ranking [14], and traditional PageRank(PR). The parameters are set through a large number of experiments and applied with $\alpha = 1800\,\text{s}$ in Eq. 1 and $d = 0.2$ in Eq. 15.

5.1 Dataset

We collected burst topic dataset from CLEar system. The system can detect and summarize burst topics in Singapore Twitter stream as soon as they emerge in real-time, which is convenient for us to collect burst topic features, tweet data and users data involved in burst topics. The collected burst topic dataset covered the period from November 1 to November 30 in 2015. Furthermore, in order to conduct burst topic user graph, the follower/followee relationships of burst topic users were also collected.

5.2 Influenced Followers Ratio

In this section, we compare the influence of the top users in each ranking approach. To evaluate this, we create a simple indicator called Influenced Followers Ratio for a burst topic k, IFR_k, defined as the fraction of followers of top N users in burst

topic that post the tweets related to burst topic k. In the three ranking approaches, the value of N is determined by PM ranking, which is given by:

$$N = \sum N_l (0 < l \leq |C_l|) \tag{16}$$

Table 1 shows the average Influenced Followers ratio (IFR) of PM ranking, TS ranking and traditional PageRank (PR) in our dataset. As shown in Table 1, Influenced Followers Ratio in PM is bigger than TS and PR, which shows that top users in PM ranking influence more their followers to spread burst topics than other ranking approaches.

Table 1. Influenced followers ratio

Approaches	IFR
PM	0.141
TS	0.098
PR	0.071

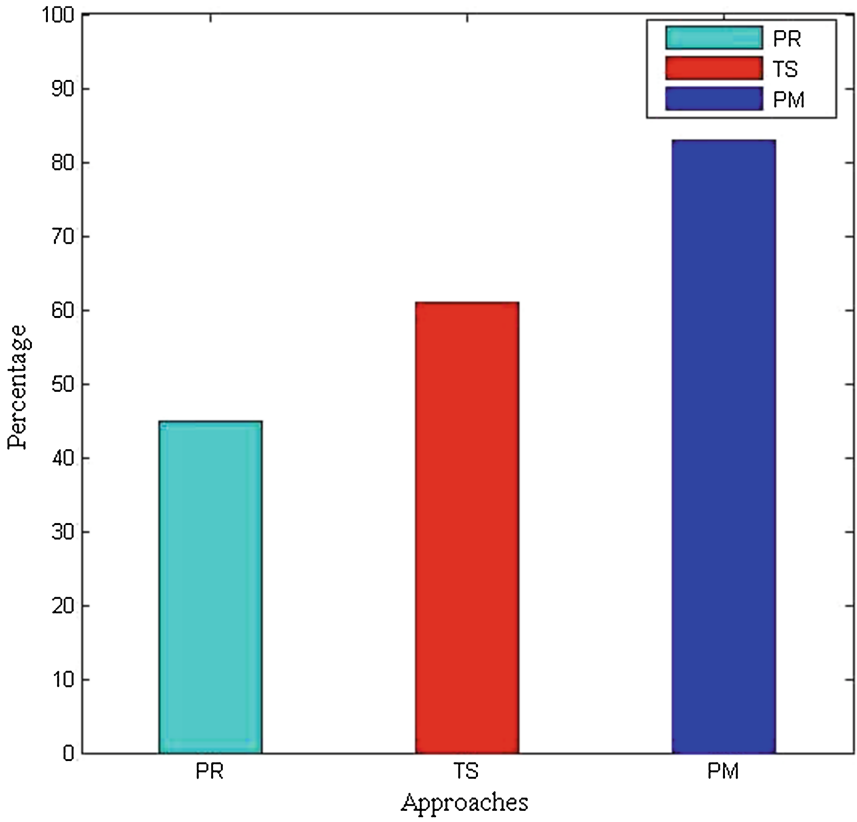

Fig. 3. The percentage of top users of each ranking that participate in the burst topic before the burst

5.3 Promoting Early Diffusion in the Early Stages

To compare the ability of top users in promoting early diffusion in the early stages of burst topic, we first obtained the detecting time of each burst in burst topic in our dataset. Due to different burst patterns, we formalize the median of detecting times of each burst in burst topic k as B_k. Next, we have to compare it with the time of top N users of each ranking that participates in the burst topic, where $T_k(r)$ represents the participation time of the user that rank r in burst topic k. If $B_k - T_k(r) < 0$, this means that the user participates in the burst topic before the burst. Finally, in our burst topic dataset, we compute the percentage of top N users of each ranking that participates in the burst topic before the burst, which is shown in Fig. 3. As shown in Fig. 3, the percentage of top users of PM ranking is larger than other approaches. More than 80 % of the top users participate in the burst topic before the burst, which indicates that our approach is more effective to detect the users that promote early diffusion in the early stages of burst topic.

6 Conclusions

In this paper, we proposed the problem of detecting community pacemakers in burst topics. In order to represent the topology structure of burst topic propagation across a large number of Twitter users, a burst topic user graph model is proposed. On one hand, a community pacemakers detection algorithm is proposed to detect community pacemakers in each large-scale user community of burst topic. On the other hand, we implement DCPBT framework on CLEar system, which can demonstrate the effectiveness of DCPBT framework. Experimental results show that our method is more effective to detect the users that influence other users and promote early diffusion in the early stages of burst topic.

Acknowledgment. This work is supported by the International Exchange Program of Harbin Engineering University for Innovation-oriented Talents Cultivation, China Scholarship Council, the Fundamental Research Funds for the Central Universities (no. HEUCF100605), the National High Technology Research and Development Program of China (no. 2012AA012802) and the National Natural Science Foundation of China (no. 61170242, no. 61572459); the National Research Foundation, Prime Ministers Office, Singapore under its International Research Centres in Singapore Funding Initiative and Pinnacle Lab for Analytics at Singapore Management University.

References

1. Kasiviswanathan, S.P., Melville, P., Banerjee, A., Sindhwani, V.: Emerging topic detection using dictionary learning. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 745–754. ACM (2011)
2. Agarwal, M.K., Ramamritham, K., Bhide, M.: Real time discovery of dense clusters in highly dynamic graphs: identifying real world events in highly dynamic environments. Proc. VLDB Endow. **5**(10), 980–991 (2012)

3. Alvanaki, F., Sebastian, M., Ramamritham, K., Weikum, G.: EnBlogue: emergent topic detection in Web 2.0 streams. In: Proceedings of the 2011 ACM SIGMOD International Conference on Management of data, pp. 1271–1274. ACM (2011)
4. Takahashi, T., Tomioka, R., Yamanishi, K.: Discovering emerging topics in social streams via link anomaly detection. In: IEEE 11th International Conference on Data Mining (ICDM), pp. 1230–1235. IEEE (2011)
5. Wang, Y., Liu, H., Lin, H., Wu, J., Wu, Z., Cao, J.: SEA: a system for event analysis on Chinese tweets. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1498–1501. ACM (2013)
6. Xie, W., Zhu, F., Jiang, J., Lim, E.P., Wang, K.: Topicsketch: real-time bursty topic detection from Twitter. In: IEEE 13th International Conference on Data Mining (ICDM), pp. 837–846. IEEE (2013)
7. Xie, R., Zhu, F., Ma, H., Xie, W., Lin, C.: CLEar: a real-time online observatory for bursty and viral events. Proc. VLDB Endow. **7**(13), 1637–1640 (2014)
8. Shen, G., Yang, W., Wang, W.: Burst topic detection oriented large-scale microblogs streams. J. Comput. Res. Dev. **52**(2), 512–521 (2015). (in Chinese)
9. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring user influence in Twitter: the million follower fallacy. In: Fourth International AAAI Conference on Weblogs and Social Media (ICWSM 2010), pp. 10–17. AAAI Press (2010)
10. Lee, C., Kwak, H., Park, H., Moon, S.: Finding influentials based on the temporal order of information adoption in Twitter. In: Proceedings of the 19th International Conference on World Wide Web, pp. 1137–1138. ACM (2010)
11. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, pp. 261–270. ACM (2010)
12. Brown, P.E., Feng, J.: Measuring user influence on Twitter using modified K-shell decomposition. In: Fifth International AAAI Conference on Weblogs and Social Media, pp. 18–23. AAAI Press (2011)
13. Fang, Q., Sang, J., Xu, C., Rui, Y.: Topic-sensitive influencer mining in interest-based social media networks via hypergraph learning. IEEE Trans. Multimedia **16**(3), 796–812 (2014)
14. Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R., Benevenuto, F.: Finding trendsetters in information networks. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1014–1022. ACM (2012)
15. Wu, Y., Hu, Y., He, X., Deng, K.: Impact of user influence on information multi-step communication in a microblog. Chin. Phys. B **23**(6), 5–12 (2014)
16. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on Twitter. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 65–74. ACM (2011)
17. Liu, D., Wu, Q., Han, W.: Measuring micro-blogging user influence based on user-tweet interaction model. In: Tan, Y., Shi, Y., Mo, H. (eds.) ICSI 2013, Part II. LNCS, vol. 7929, pp. 146–153. Springer, Heidelberg (2013)
18. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. **105**(4), 1118–1123 (2008)
19. Lambiotte, R., Rosvall, M.: Ranking and clustering of nodes in networks with smart teleportation. Phys. Rev. E **85**(5), 056107(1–9) (2012)