Wireless Networks (2022) 28:1169–1175
https://doi.org/10.1007/s11276-018-01913-4

User interest community detection on social media using collaborative filtering

Liang Jiang[1,3] · Leilei Shi[1] · Lu Liu[2] · Jingjing Yao[4] · Moses Edward Ali[2]

Published online: 27 February 2019

Abstract
Community detection in microblogging environment has become an important tool to understand the emerging events. Most existing community detection methods only use network topology of users to identify optimal communities. These methods ignore the structural information of the posts and the semantic information of users' interests. To overcome these challenges, this paper uses User Interest Community Detection model to analyze text streams from microblogging sites for detecting users' interest communities. We propose HITS Latent Dirichlet Allocation model based on modified Hypertext Induced Topic Search and Latent Dirichlet Allocation to distil emerging interests and high-influence users by reducing negative impact of non-related users and its interests. Moreover, we propose HITS Label Propagation Algorithm method based on Label Propagation Algorithm and Collaborative Filtering to segregate the community interests of users more accurately and efficiently. Our experimental results demonstrate the effectiveness of our model on users' interest community detection and in addressing the data sparsity problem of the posts.

Keywords Interest detection · Social network · UICD · HLDA · HLPA

1 Introduction

With the rapid developments of Internet, many different forms of online social networks (OSNs) have emerged and have attracted a wide spectrum of users. Some of the most popular sites today are Facebook, Twitter and Weibo [1]. The popularity of online social networks makes it an important platform to share and gather information at the same time. However, due to huge volume of generated data, users cannot find their interested content from heterogeneous sources. This problem is known as information overload and it is becoming serious with the continuous expansion of the scale of social networks. However, to address this problem, community detection is recommended as an effective way to solve it. For community detection, structure of the community is a primary feature that should be considered [2]. Research studies highlights that there are many algorithms on the community detection, such as, algorithms based on module optimization [3], spectral clustering [4], hierarchical classification [5], and label propagation [6]. However, these community detection methods only use user's network topology to identify optimal communities. Applications of such methods suffer from interest drift. Moreover, the limitation of characters' length in each user's post makes it more difficult to obtain the interest of users.

To overcome these challenges, user interest community detection model is proposed, named User Interest Community Detection (UICD) model. This paper uses modified Hypertext Induced Topic Search (HITS) [16] and Latent Dirichlet Allocation (LDA) [7] based interest detection model named HITS Latent Dirichlet Allocation (HLDA) to distil hot interests and high-influence users by reducing negative impact of meaningless ordinary users and interests. Then, Label Propagation Algorithm (LPA) and Collaborative Filtering Recommender based user interest community detection method named HITS Label Propa-

✉ Lu Liu
l.liu@derby.ac.uk

1 School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China

2 Department of Computing and Mathematics, University of Derby, Derby, UK

3 Jingjiang College of Jiangsu University, Zhenjiang, China

4 School of Economy and Finance, Jiangsu University, Zhenjiang, China

gation Algorithm (HLPA) is proposed to divide the interest communities accurately and efficiently.

The main contributions of this paper are listed as follows:

1. We propose modified HITS algorithm-based user interest filtering and LDA based interest detection model to distil emerging interests and high-influence users by reducing negative impact of non-related users and its interests. Finally, the nearest neighbor set of the target user is formed, and the result is recommended according to the user rating data of the nearest neighbor.
2. We proposed LPA and Collaborative Filtering Recommender based user interest community detection method named HLPA. The proposed method assigns a unique tag to each post, and then updates the post's label in the order of high to low. Finally, stable user interest communities can be obtained when the update is finished.
3. Experimental results on real-world networks indicate that the UICD model provides richer information for the inner structure of the detected users' interest communities. With the help of prototype weights compared with the existing community detection models, and experiment results indicate the efficiency of our approach at the same time.

The rest of this paper is structured as follows: in Sect. 2, we discuss related work for detecting users' interest community in microblogging networks. In Sect. 3, we introduce our models concerning community detection. The results of our experiments are presented in Sect. 4. The last section concludes and provides future study.

2 Related work

Community detection is becoming a very popular research field, especially within the domain of social media [8]. It has attracted many researchers due to its openness and the availability of data. Twitter access through Twitter API and Facebook access through Facebook API is widely used by developers these days. It is found that the phenomenon of community in many social networks is ubiquitous [9]. It contains various large and small, explicit and implicit communities. This kind of network structure has one common feature, that the nodes have a very close relationship with the community [10], [11].

There are two main types of classification algorithms for the social network: (1) To divide the community by the dichotomy of graph theory. (2) To divide the community by the idea of clustering algorithm. Modern typical community classification algorithms are mainly scattered. There are some existing researches based on module optimization algorithm. Cao et al. [3] proposed a new community detection method, to find crisp and fuzzy communities in undirected and unweighted networks by maximizing weighted modularity. Xin et al. [12] proposed the RWS (Random Walk Sampling) method to detect the overlapping communities, utilizing the random walk method to find the closest friends for each node.

Some existing researches are also based on label propagation algorithm. Li et al. [6] proposed a modified LPA called Stepping LPA-S, in which labels are propagated by similarity. Peng et al. [13] proposed an improved label propagation algorithm (LPA) to uncover overlapping community structure. However, the LPA algorithm uses the random selection scheme, which makes the result of the community detection unstable, and the quality of the community is also very difficult to be guaranteed. Inspired by this, Cao et al. [14] proposed the first community-based influence maximization algorithm OASNET (Optimal Allocation in a Social Network). However, the community detection methods mentioned above only use users' network topology to identify optimal communities. Direct application of such methods for interest community detection suffers from interest drift and data sparsity. Moreover, the UICD model can solve the problem of data sparsity of posts' information.

3 The user interest community detection model

The HLDA model is composed of two modules: (1) the scored interests according to the HITS method. Posts in the network are clustered by LDA topic model and Gibbs sampling [7], [15]. (2) the scored interests according to the HITS method [16], [17]. The most influential users among emerging interests are discovered through authority value in the interest-user network. Figure 1 illustrates the process involved in interest detection and discovery of the influential users in hot interests.

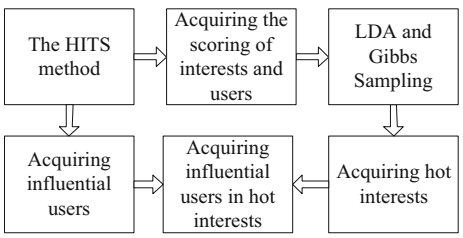

Fig. 1 The procedure of the HLDA model

3.1 Extracting hot interests and high-quality users based on HITS algorithm

In this paper, HITS algorithm is extended to exploit the inseparable connection between the interests and their corresponding users for distilling the high-quality users and the popular interests. The proposed Filtering User-Interest method can effectively filter random low-quality users and ordinary interests.

Many hot interests generally attract a high-quality user. Intuitively, hot interests can be recommended or commented by more number of high-quality users as compared to that of the ordinary interests. In addition, high-quality users can draw an increased level of attention to the hot interests, which are usually spread or broadcast over the microblogging network. The authority value of the HITS algorithm has been completely utilized with more importance, to identify the high-quality users, alongside the hub value of the interests. Furthermore, a special emphasis has been given to the theory that there is a strong possibility of hot interests attracting many high-quality users.

3.2 Fusion of user similarity and user trust degree

User trust degree calculation Definition 1 (social network S) Let U denote user nodes set in set S, I denote project set in set S, $E1$ denote a direct trust relationship between users in set S, $E2$ denote evaluation relationship set between user and project in set S, $W1$ denote direct trust set T among users in set S, $W2$ denote user rating project set R in set S, Then S can be expressed as a six-tuple $S(U, I, E1, E2, W1, W2)$. Where $U = \{u1, u2, \ldots, un\}$, $|U| = n$; $I = \{i1, i2, \ldots, im\}$, $|I| = m$; $E1 = \{\langle u, v\rangle | u, v \in U\}$; $E2 = \{\langle u, i\rangle | u \in U, i \in I\}$; $W1 = \{T(u, v) | u, v \in U\}$; $W2 = \{R(u, i) | u \in U, i \in I \bigwedge \langle u, i\rangle \in E2\}$.

User network definition By definition 1 social network $S(U, I, E_1, E_2, W_1, W_2)$ can extract two sub-networks. Namely the user network $G(U', E', W')$ and user-project network $G(U'', I', E'', W'')$. Where the user network $G(U', E', W')$ is defined in Definition 2.

Definition 2 (user network G) G denotes the relationship network made by contact between users, and represent by $G(U', E', W')$. Where $G \subset S$; U' denotes the set of user nodes, $U' \subseteq U$; E' denotes the set of direct trust relationship between users, $E' \subseteq E_1$, $E' = \{\langle u, v\rangle | u, v \in U'\}$; W' denotes a collection of direct trust degree T among users, $W' \subseteq W_1$, $W' = \{T(u, v) = 1 | u, v \in U' \wedge \langle u, v\rangle \in E'\}$.

Direct trust degree calculation Definition 3 (direct trust degree T) If there is the edge from user node u to user node v in user network G, the direct trust degree $T(u, v)$ of u to v is 1, otherwise is 0.

Considering the value range of the selected user similarity measure method is [− 1,1], in order to integrate user trust degree and similarity of users conveniently, so it is necessary to normalize direct trust degree T making the normalized direct trust degree tr value is limited in the range of [0, 1].

The direct trust degree obtained after the normalization process is shown in the formula (1).

$$tr(u, v) = \frac{T(u, v)}{\sum_{u' \in F(u)} T(u, u')} \tag{1}$$

where $F(u) = \{u' | \langle u, u'\rangle \in E' \wedge u, u' \in U'\}$, $tr(u, v)$ denote normalized direct trust degree of u to v, and tr(u, v) satisfies $\sum_{v \in F(u)} tr(u, v) = 1$.

Indirect trust degree calculation Definition 4 (indirect trust degree T) If there is at least one path from user node u to user node v in user network G, It can be considered that u has an indirect trust relationship with v, and the shortest path from u to v is $path = \{\langle u, a_0, a_1, \ldots, a_k, v\rangle | \min(k + 1) \wedge (k + 1) > 1, u, a_x, v \in U', 0 \leq x \leq k\}$, then the indirect trust degree of u to v is $T'(u, v) = (tr(u, a_0) + tr(a_0, a_1) + \cdots + tr(a_k, v))/(k + 1)$.

In addition, according to the small world theory set $max(k + 1)$ to 6. If $min(k + 1)$ of user node u to user node v in user network G is greater than 6, $T'(u, v) = 0$.

User trust degree calculation User trust is an important factor that affects the personalized recommendation of social networks. As mentioned above, the user trust is divided into direct trust and indirect trust, so the user trust is their comprehensive value.

Let $Tr(u, v)$ denotes that user trust degree of user node u to user node v, $tr(u, v)$ denotes that normalized direct trust degree of user node u to user node v, $T'(u, v)$ denotes that indirect trust degree of user node u to user node v.

If $tr(u, v)$ is expressed by A, $T'(u, v)$ is expressed by B, the user's trust degree is calculated as formula (2).

$$Tr(u, v) = \begin{cases} \frac{2 \times AB}{A + B} & A > 0 \wedge B > 0 \\ A & A > 0 \wedge B = 0 \\ B & A = 0 \wedge B > 0 \end{cases} \tag{2}$$

User similarity calculation User similarity is calculated to find users with similar interests and preferences, thereby generating recommendation results. User similarity calculation of u and v on user network G is shown in formula (3).

$$Si(u,v) = \frac{\sum_{i=1}^{m}\left(R_{u,i} - \overline{R_u}\right)\left(R_{v,i} - \overline{R_v}\right)}{\sqrt{\sum_{i=1}^{m}\left(R_{u,i} - \overline{R_u}\right)^2}\sqrt{\sum_{i=1}^{m}\left(R_{v,i} - \overline{R_v}\right)^2}} \quad (3)$$

where $R_{u,I}$ and $R_{v,i}$ respectively denote the user u and user v ratings for project i. $\overline{R_u}$ and $\overline{R_v}$ respectively denote the user u and user v the average ratings for all projects.

Fusion of user similarity and user trust degree The user similarity and user trust degree are fused together to alleviate the problem of sparse data.

Let $Si(u, v)$ denotes user similarity of user u and user v, $Tr(u, v)$ denotes user similarity of user u to user v. $We(u, v)$ denotes a comprehensive value after fusing $Si(u, v)$ and $Tr(u, v)$, then the calculation of $We(u, v)$ is shown as formula (4).

$$We(u,v) = \lambda \times Si(u,v) + \mu \times Tr(u,v) \quad (4)$$

where λ, μ respectively denote the proportion for $Si(u, v)$, $Tr(u, v)$, and $\lambda + \mu$=1.

3.3 HLPA model

In 2007, Raghavan et al. [18] proposed the classification algorithm based on label propagation. This is the first time in the domain of the label spreading. So many scholars referred to the algorithm as the LPA algorithm.

The complexity of LPA algorithm is almost linear time and the design of algorithm is very simple. This makes the LPA algorithm widely discussed among many scholars. However, the LPA algorithm uses the random selection scheme, which makes the result very unstable, and the quality of the community is difficult to be guaranteed. For this purpose, the stability of the LPA algorithm is improved. A stable and high quality method based on node influence is proposed, named HLPA.

Figure 2 illustrates the process involved in user interest community detection. Specifically, the HLPA algorithm assigns a unique tag to each node, and then updates the node's label in the order of high to low. In each iteration of the label updates, selecting majority of the adjacent nodes are held by the label to be updated. If there is more than one tag, then calculate the influence of the label and select the largest influence from the updates. The update is finished and can be obtained by a stable community. Finally, the stability and quality of HLPA algorithm are verified on the real data set. However, the LPA algorithm uses the random selection scheme, which makes the final result of the community very unstable, and the quality of the community is also very difficult to be guaranteed. If the selection of the community is lower, it means that the relationship between the nodes in the community is not enough. When the network is divided into an independent community, the link between the community will be cut off, so once the link between the nodes in the community, the impact of the node will be more, the impact of the node will be more inaccurate, and then lead to the final selection of the so-called core node set will be greatly reduced. On the contrary, if there is a scheme to ensure that each of the selected community division results are high quality, then the impact of the core nodes selected from various communities will be guaranteed.

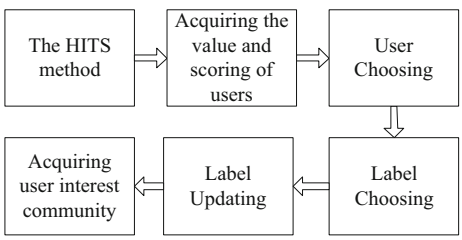

Fig. 2 The procedure of the HLPA method

4 Experiments

In this section the details of our experiments, conducted on real-world collected data, are demonstrated. It reflects the effectiveness of our proposed UICD method. It further describes the data collection and processing steps, environment setup and analysis, comparative methods comparison, evaluation criteria and result analysis.

4.1 Dataset

Our dataset are collected from Twitter (http://twitter.com/) via Twitter API. The collected dataset is composed of 1,000,000 posts from April 25, 2018 to April 28, 2018. As discussed earlier, to reduce the impact of the bump phenomenon, only the users who publish or comment on posts are considered in our dataset.

4.2 Experimental settings

The experiments were conducted on a machine with Intel I7 4.2 GHz CPU and 16G memory.

As shown in Fig. 3, more than 99% of the total number of labels lies into the range of 1–6. Hence, while comparing with other methods, our work only recommends labels within 1–6 ranges. In this paper, 7145 labels are divided into five equal parts and the five-fold cross validation is used to evaluate the proposed method with other methods. The average of the five assessment results is obtained. In each evaluation, parameters were tuned via grid search. For LDA, $\alpha = 0.5$ and $\beta = 0.1$. In all the methods, Gibbs sampling was run for 1000 iterations. The results reported

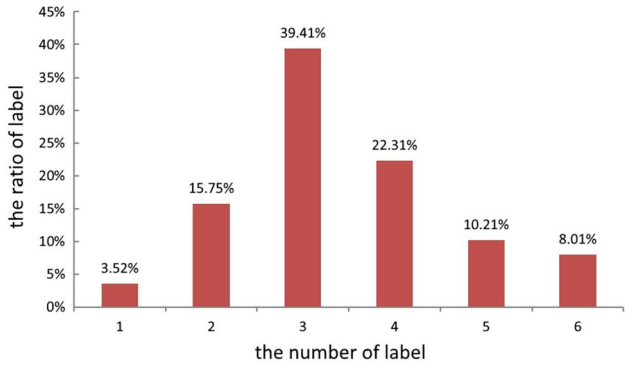

Fig. 3 The distribution of 7145 labels

are captured on average after 5 runs. To filter high-quality posts, all initial authority scores *d.a* and *u.h* hub scores were set to 1. For fusion of user similarity and user trust degree, $\lambda = 0.72$ and $\mu = 0.28$, according to the experiments.

4.3 Comparative methods and evaluation criteria

For comparative analysis, two recommended methods are used, based on TF-IDF and LDA. The following evaluation criteria are used: Recall, Precision and F-measure. They are defined as follows:

$$Recall = \frac{|labels_r \cap labels_m|}{|labels_m|},$$
$$Precision = \frac{|labels_r \cap labels_m|}{|labels_r|} \qquad (5, 6, 7)$$
$$F{-}measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

where $labels_r$ represents a recommended set of labels, $labels_m$ represents a true set of labels.

4.4 Result analysis

Figure 4 shows the recommended effect of our proposed method, Collaborative Filtering Recommender based User Interest Community Detection. The results obtained are compared with TF-IDF and LDA methods. It is evident from the figures, on the basis of our adopted evaluation criteria that the recommended effect of the method based on topic model is better than that of the keyword matching based method. This difference is because of the two basic characteristics of the label: the arbitrariness of the user's markup and the semantic ambiguity of the label. Both of these characteristics hinder the recommendation system to recommend relevant and accurate labels. The label recommendation method based on keyword matching technology only uses statistical information between the documents, which might lose some useful information, such as hidden topic information of the documents. The method based on semantic level takes full advantage of the implicit topic information of the document. In the label recommendation system, it is significant to make full use of the implicit topic information of the document to improve the recommendation effect.

Furthermore, as seen from the figures, our proposed method has better recall, precision and F-measure than the other two methods when recommending labels. As the number of recommended labels increases gradually, the recall rate of all methods is increasing because of the increasing number of correct labels. However, as the number of real labels is limited, it can be predicted that the recall rate would be steady with the increase of the recommended labels. On the contrary, the precision rate of all methods decreases as the number of labels increases. This is because more inaccurate labels are recommended. The average F-measure value of the proposed method is 10.8% and 21.7% higher than that of LDA and TF-IDF respectively when 1 to 6 labels are recommended.

Figure 5a shows comparison results of sorting and non-sorting interest set, the abscissa represents the number of recommended interests, and the ordinate represents the F-measure. This paper uses interest detection model named HLDA to distil hot interests and high-influence users. At the same time, the interests are sorted according to the popularity of interest. As can be seen from Fig. 5a, sorting the recommended interest set will significantly improve the accuracy of the recommended interests. This is because

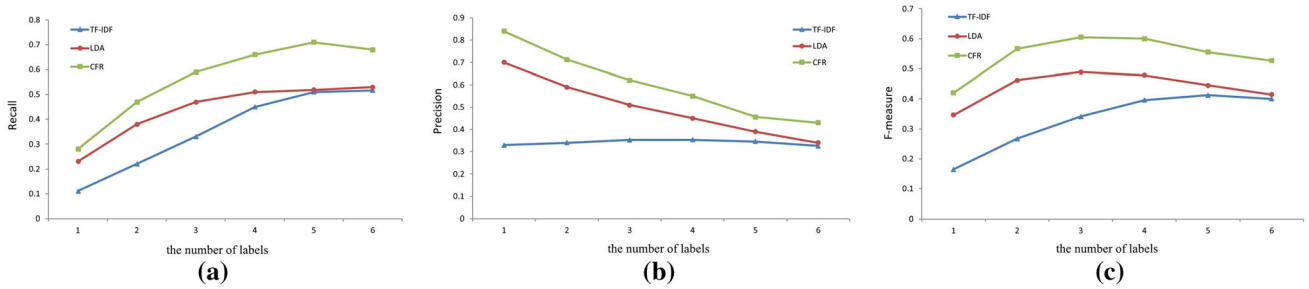

Fig. 4 Recall, precision and F-measure rate comparison

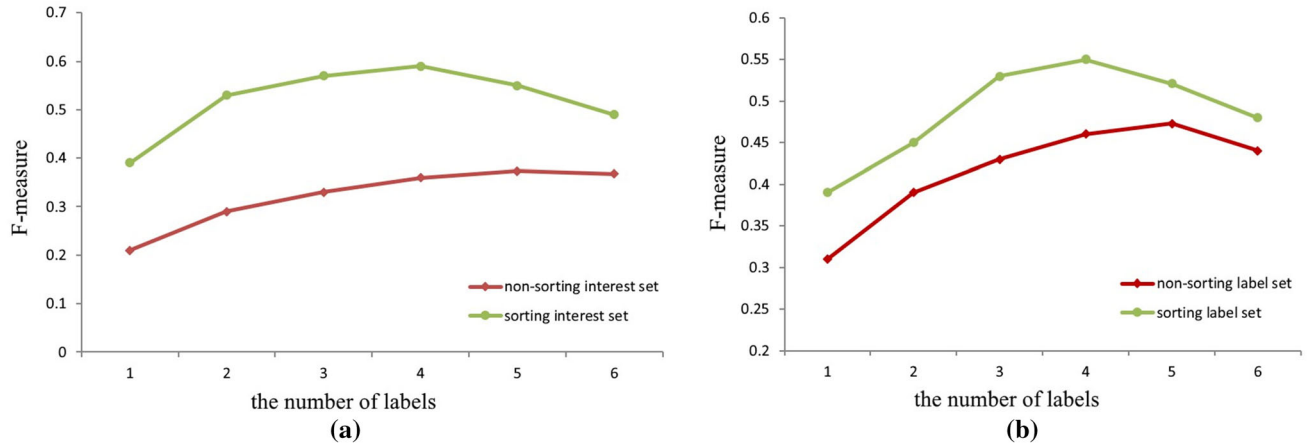

Fig. 5 Comparison results of sorting and non-sorting interest/label set

there are large amounts of meaningless ordinary interests in the interest set, if the interest set is used directly for the recommendation, it cannot achieve good results. The interests with high popularity are given priority to recommend by sorting.

Figure 5b shows comparison results of sorting and non-sorting of label set, the abscissa represents the number of recommended labels, and the ordinate represents the F-measure. The LPA algorithm uses the random selection scheme, which makes the result very unstable, and the quality of the community is difficult to be guaranteed. To improve the LPA algorithm, we propose a stable and high quality method based on node influence named HLPA. The HLPA algorithm assigns a unique tag to each node, and then updates the node's label in the order of high to low. In each iteration of the label updates, selecting majority of the adjacent nodes are held by the label to be updated. If there is more than one tag, then calculate the influence of the label and select the largest influence from the updates. The update is finished and can be obtained by a stable community.

5 Conclusion and future work

The popularity of social networks makes it an important platform for people to share information. In this paper, we successfully focused on users' interest community detection by analyzing the text stream of the microblogging sites using our proposed model. Our work used modified HITS and LDA based interest detection model named HLDA to distill emerging interests and high-influence users. Furthermore, LPA and Collaborative Filtering Recommender based user interest community detection method named HLPA is proposed to segregate the interest community of users more accurately and efficiently. For comparative analysis, two methods based on TF-IDF and LDA are used. Evaluation criteria such as recall, precision and F-measure are used and the obtained experimental results demonstrated the effectiveness of our proposed model.

Acknowledgements This work was partially supported by the National Natural Science Foundation of China under Grants Nos. 61502209, 61502207 and 71701082, Natural Science Foundation of Jiangsu Province under Grant BK20170069, UK-Jiangsu 20-20 World Class University Initiative programme, UK-China Knowledge Economy Education Partnership and Postgraduate Research & Practice Innovation Program of Jiangsu Province No. KYCX17_1808.

References

1. Gao, Q., Abel, F., Houben, G. J. et al. (2012). A comparative study of users' microblogging behavior on Sina Weibo and Twitter. In *User Modeling, adaptation, and personalization* (pp. 88-101). Springer, Berlin.
2. Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486,* 75–174.
3. Cao, J., Bu, Z., Gao, G., & Tao, H. (2016). Weighted modularity optimization for crisp and fuzzy community detection in large-scale networks. *Physica A: Statistical Mechanics and its Applications, 462,* 386–395.
4. Shen, G., & Ye, D. (2017). A distance-based spectral clustering approach with applications to network community detection. *Journal of Industrial Information Integration, 6,* 22–32.
5. Raj, E. D., & Babu, L. D. D. (2016). A fuzzy adaptive resonance theory inspired overlapping community detection method for online social networks. *Knowledge-Based Systems, 113,* 75–87.
6. Li, W., Huang, C., Wang, M., & Chen, X. (2017). Stepping community detection algorithm based on label propagation and similarity. *Physica A: Statistical Mechanics & Its Applications, 472,* 145–155.

7. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.
8. Hwang, W. S., Li, S., Kim, S. W., & Lee, K. (2014). *Data imputation using a trust network for recommendation* (pp. 299–300).
9. Yin, C., & Chu, T. (2013). Improving personal product recommendation via friendships' expansion. *Journal of Computer & Communications, 01,* 1–8.
10. Deng, S. G., Huang, L. T., Wu, J., & Wu, Z. H. (2015). Trust-based personalized service recommendation: A network perspective. In *IEEE international conference on multimedia and expo* (pp. 1–6).
11. Yuan, B., Liu, L., & Antonopoulos, N. (2018). Efficient service discovery in decentralized online social networks. *Future Generation Computer Systems, 88*, 775–791.
12. Xin, Y., Xie, Z. Q., & Yang, J. (2016). *An adaptive random walk sampling method on dynamic community detection*. Oxford: Pergamon Press, Inc.
13. Peng, H., Zhao, D., Li, L., Lu, J., Han, J., & Wu, S. (2016). An improved label propagation algorithm using average node energy in complex networks. *Physica A: Statistical Mechanics and its Applications, 460,* 98–104.
14. Cao, T., Wu, X., Wang, S., & Hu, X.: OASNET: An optimal allocation approach to influence maximization in modular social networks. In *ACM symposium on applied computing* (pp. 1088–1094).
15. Guo, Y., Liu, L., Wu, Y., & Hardy, J. (2018). Interest-aware content discovery in peer-to-peer social networks. *ACM Transactions on Internet Technology, 18*(3), 1–21.
16. Shi, L. L., Liu, L., Wu, Y., Jiang, L., & Hardy, J. (2017). Event detection and user interest discovering in social media data streams. *IEEE Access, 5*(99), 20953–20964.
17. Sun, X., Wu, Y., Liu, L., & Panneerselvam, J. (2015). Efficient event detection in social media data streams. In *IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing* (pp. 1711–1717).
18. Raghavan, U. N., Albert, R., & Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics, 76,* 036106.

Liang Jiang received the B.S. degree from the Nanjing University of Posts and Telecommunications, China, in 2007, and the M.S. degree from Jiangsu University, Zhenjiang, China, in 2011, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Telecommunication Engineering. His research interests include OSNs, computer networks, and network security.

Leilei Shi received the B.S. degree from Nantong University, Nantong, China, in 2012, and the M.S. degree from Jiangsu University, Zhenjiang, China, in 2015, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Telecommunication Engineering. His research interests include event detection, data mining, social computing, and cloud computing.

Lu Liu received the M.S. degree from Brunel University and the Ph.D. degree from the University of Surrey. He is currently a Professor of Distributed Computing with the University of Derby, U.K, and an Adjunct Professor with Jiangsu University, China. His research interests are in areas of cloud computing, social computing, service-oriented computing, and peer-to-peer computing. He is a fellow of the British Computer Society.

Jingjing Yao received the B.E. degree from Jiangsu University, Zhenjiang, China, in 2011, and the D.M. degree from Jiangsu University, Zhenjiang, China, in 2016. Her research interests include complex network, information dissemination.

Moses Edward Ali received the B.Sc. in Computer Engineer and the M.Sc. in Software Engineering from University of Southampton, UK. He is a Ph.D. research student in the Department of Electronics, Computing & Mathematics, College of Engineering & Technology at the University of Derby, UK. His research interests lie in the area of Social Media Analyses, Machine Learning on real-time data streams, Semantic Web and Innovative Software Development.