Probabilistic model based large-scale social network community discovery algorithm

Xu Dong-fang

(Basic Courses Department, Henan Polytechnic; Henan Zhengzhou 450000,China)

Tian Chang-shen

(Henan Polytechnic; Henan Zhengzhou 450000,China)
xudongfang2014_cn@126.com

***Abstract*—In this paper, we propose a novel large-scale social network community discovery algorithm based on probabilistic model to organize users with similar interests into a same group. Firstly, the user community discovery problem is illustrated. The large-scale social network can be regarded as a graph, in which edge represents the relationship between two nodes. Therefore, the user community detection problem can be converted to the graph partition problem. Secondly, our proposed user community discovery algorithm is given. Our algorithm follows an assumption that users of a same community are possible to have same or similar interests. Therefore, the main innovations of our algorithm lie in that the community topics are be represented as multinomial distribution on words, and user interests in different topics obey the probabilistic distribution on community topics. Finally, experiments are conducted to make performance evolution. Experimental results demonstrate that our proposed algorithm can effectively solve the problem of user community detection for the large-scale social network than other methods.**

Keywords- Probabilistic model, Community discovery, Large-scale social network, Multinomial distribution

I. INTRODUCTION

In recent years, many online social networks have developed rapidly based on the Web 2.0 platform, such as Facebook, Twitter, Myspace, and so on. Particularly, a social network contains several different types of relations among different social actors. For example, on Twitter, a user can specify a person to follow, and then make up an explicit friendship network. Meanwhile, this user's posted tweets can provide important clues about her interests. Furthermore, users with same or similar interests can be classified into a same group, this is, user community[1][2].

Online social network is belonged to the complex network, which is a powerful method to understand the complex systems, which have attracted more and more attentions of researchers. In the former studies, researchers have found that social networks have several statistical properties, such as small world property, scale free distribution, and so on. Among these properties, community structure is of great importance. In a community structure, the nodes are classified into several groups within which the network connections are dense, however, between which the networks connection are sparser[3-6]. However, the precise definition of community is difficult to make, the reason lies in that in many cases communities may overlap with each other, and each node may occur in at least one communities.

For the social networks, user community detecting refers to the process of seeking dense groups on graphs such as users with same or similar interests. Furthermore, community members have more common properties amongst themselves than others in the community, and the identification of community structure can provide an effective way to analyze the functionalities of networks.

Based on the above analysis, in this paper, we focus on the problem of detect the user community in large-scale social network, and a probabilistic model are utilized. Probabilistic model has been proved to be an effective way to solve many complex computing problem[7-10]. The rest of the paper is organized as follows. Section 2 illustrates the user community discovery problem. In section 3, the user community discovery algorithm based on the probabilistic model is presented. Section 4 proposes experimental results and provides related analysis. Finally, the conclusions are drawn in section 5.

II. ILLUSTRATION OF THE USER COMMUNITY DISCOVERY PROBLEM

978-1-4799-6636-3/14 $31.00 © 2014 IEEE
DOI 10.1109/ICICTA.2014.110

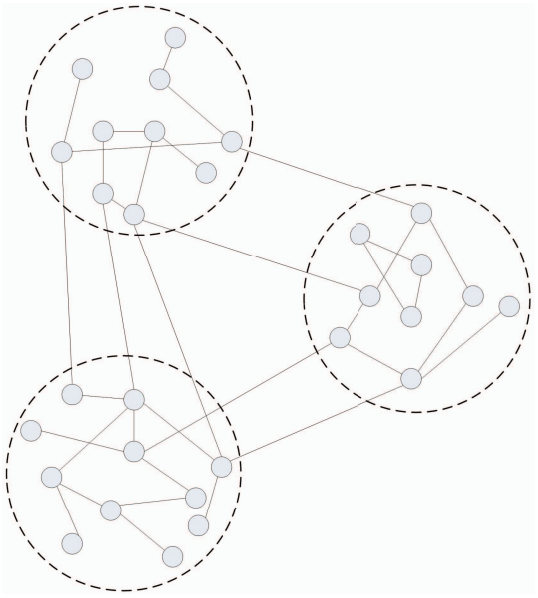

Figure 1. Example of a network with community structure

The large-scale social network can be regarded as a graph $G(V,E)$ with the set of vertices $V=(v_1,v_2,\cdots,v_n)$, in which edge represents the relationship between two nodes. Hence, the user community detection approach aims to obtain the node partition $P=C_1\bigcup C_2\bigcup\cdots\bigcup C_m$, where $C_1,C_2,\cdots,C_m$ denote the sets of vertices (also named communities) which may be disjointed or overlapped with each other. As is shown in Fig.1, an example of a network with three communities is illustrated, which have more dense internal links than the external links.Afterwards, some important definitions are given as follows.

Definition 1 (User community) User community (denoted as C) refers to a group of users who construct densen friendship connections when they attend a particularly topic. C can be computed using the following equation.

$$C=\langle \psi_C,\zeta_C\rangle \tag{1}$$

where ψ_C and ζ_C represent the community distribution for different users and the community topic distribution for different words.

Definition 2 (Community structure) The community structure of a social network N is represented as $P=\langle C_1,C_2,\cdots,C_m\rangle=\langle \psi,\zeta\rangle$, where m means the number of user communities.

III. USER COMMUNITY DISCOVERY ALGORITHM BASED ON THE PROBABILISTIC MODEL

In this section, a novel probabilistic model is described in advance, which can solve the computing problem containing "users", "words", "topics" and "communities". Next, some important symbols are described in Table.1

Table.1 Symbols description

α	Community distribution set for the user set
α_u	Community distribution for the user u
θ	Word distribution set for topic set
θ_z	Word distribution set for topic z
μ	Topic distribution set for community set
μ_c	Topic distribution set for community c
υ	User distribution set for community set
υ_c	User distribution set for community c

Based on the above symbols descriptions, for a specific user u in a social network, the generative process of our probabilistic model is given as follows.

(1) Generating each word for the user u

1) Sampling a community c from the multinomial distribution α_u.

2) Sampling a topic z from the multinomial distribution μ_c.

3) Sampling a word w from the multinomial distribution θ_z.

(2) Generating each edge of graph for the user u.

1) Sampling a community c from the multinomial distribution α_u

2) Sampling a user v from the multinomial distribution υ_c and construct a edge from u to v.

Our proposed algorithm obeys an assumption that users belonged to a same community may share same or similar interests. Hence, to implement the user community discovery process, the community topics can be represented as multinomial distribution on words (denoted as Φ), and user interests in different topics is a probabilistic distribution on community topics (denoted as Θ).

Then, the probability of the corpus contents conditioned on Φ and Θ can be estimated as follows.

$$P(w,\Big|\Phi,\Theta,U)=\prod_{d=1}^{D}\prod_{i=1}^{N_d}\frac{1}{|U_d|}\sum_{a\in U_d}\sum_{k=1}^{r}\Theta_{w_{i,u}k}\cdot\Phi_{ka} \tag{2}$$

where N_d refers the sampling times to construct document d, and U_d represents the user set related to the document d. Particularly, the posterior distribution is evaluated utilizing the Gibbs sampling algorithm without estimating the parameters of this model. Afterwards, the community topic and the user allocation for each word can be sampled using the following equation.

$$p(z_i = k, x_i = u \mid w_i, z_i, x_i, \alpha, \beta, U) \propto \Phi_{wk}\Theta_{uk} = \frac{C_{wk}^{w\gamma} + \beta}{\sum_{w} C_{wk}^{w\gamma} + |U|\beta} \cdot \frac{C_{ku}^{\gamma U} + \alpha}{\sum_{u \in U} C_{ku}^{\gamma U} + |U|\alpha} \quad (3)$$

where $z_i = k, x_i = u$ means than the i^{th} word is allocated to the k^{th} topic and u^{th} user respectively, and α and β refer to the prior parameters for the Dirichlet distribution. Particularly, Φ and Θ can be estimated by Eq.4 and Eq.5.

$$\Phi_{wk} = \frac{C_{wk}^{w\gamma} + \beta}{\sum_{w} C_{wk}^{w\gamma} + |U|\beta} \quad (4)$$

$$\Theta_{uk} = \frac{C_{uk}^{U\gamma} + \alpha}{\sum_{u \in U} C_{uk}^{U\gamma} + |U|\alpha} \quad (5)$$

Our algorithm aims to process the user-supplied information. Through analyzing the relationship between users and topics, our algorithm can cluster users with similar interests into the same community and represent the community topic with several words.

IV. EXPERIMENT

In this section, experiments are conducted to make performance evaluation, and datasets we utilized are introduced as follows.

A. Datasets

Two datasets are used in this experiment, dataset 1 is collected from Twitter, which is a microblogging site where users can post text of up to 140 characters on their profile pages. In this dataset, 41.7 million user profiles, 1.47 billion social relations, and 106 million tweets[20]. Particularly, this dataset contains 1023 users, 5361 unique terms and 350929 links.

Dataset 2 is collected from Delicious dataset[21], and Delicious is a popular social tagging system, allowing users to bookmark various web urls with individually selected tags. To reduce the size of the original dataset, we delete users who did not post any urls and urls that had received less than five tags. After this process, this dataset is made up of 749 users.

B. Performance evaluation metric

(1) Mean value

Mean value is defined based on the soft modularity and user-content similarity to estimate the quality of the communities detected.

(2) Accuracy

For a specific user u, his label l_u in the dataset and the assigned label a_u is extracted using the given approach. Then, accuracy is defined as the following equation.

$$Accuracy = \frac{1}{|U|} \cdot \sum_{u \in U} \upsilon(l_u, m(a_u)) \quad (6)$$

where $|U|$ denotes the number of all the users and function $\upsilon(x, y)$ represents the function the value of which is equal to 1 when $x = y$, otherwise 0. $m(a_u)$ refers to a permutation mapping function.

(3) Normalized Mutual Information

Normalized mutual information is defined as follows.

$$NMI(L, \bar{L}) = \frac{MI(L, \bar{L})}{\max\left(H(L), H(\bar{L})\right)} \quad (7)$$

where L means the user labels obtained from the dataset and $\bar{L}$ denotes the labels extracted by the specific approach. $H(L)$ represents the entropy of L and $MI()$ refers to the mutual information metric.

C. Experimental results and analysis

To make performance comparison, other four typical methods are chosen, which are 1) Normalized cut[22], 2) SSNLDA[23], 3) NMF based method[24] and 4) AT model[25]. Normalized cut can be used in the graph partition and it is suitable to solve the user community detection problem. SSNLDA refers to a LDA-based hierarchical Bayesian algorithm on a link graph where communities are modeled as latent variables and defined as distributions over user space. NMF based method is design to discover fuzzy community structures in complex networks based on non-negative matrix factorization (NMF). The AT model represents the author-topic (AT) model to explore the relationships among users, documents, topics, and words. It represents a topic as a multinomial distribution over words and models a user as probability distribution over different topics.

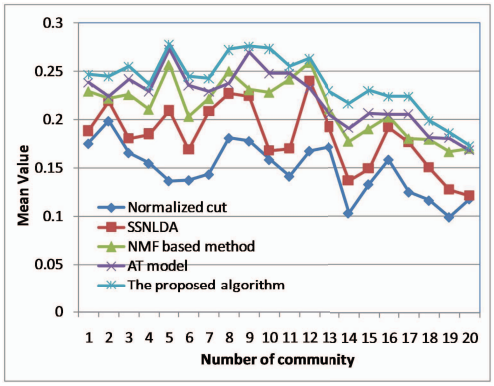

Figure 2. Mean value for different method when community number changing for dataset 1.

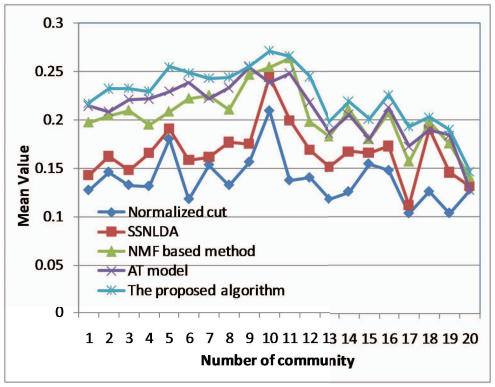

Figure 3. Mean value for different method when community number changing for dataset 2.

Fig.2 and Fig.3 show that the proposed algorithm performs better than other methods for both dataset 1 and dataset 2 for all the communities with the number changing from 1 to 20. Afterwards, the performance evaluation metric Accuracy and Normalized Mutual Information are utilized, and the experimental results are given in Table.2 and Table.3.

As is shown in Table.2 and Table.3, the proposed algorithm performs best for Accuracy and Normalized Mutual Information, and the highest value is represent by bold.

Table.2 Accuracy comparision for different methods.

Dataset	N	NC	SSNLDA	NMF	AT	The proposed algorithm
Dataset 1	5	49.69	51.24	57.12	64.85	**75.23**
	10	50.05	50.77	55.29	58.13	**63.92**
	15	39.19	43.01	43.58	46.93	**48.16**
	20	29.44	33.34	38.23	39.57	**39.85**
Dataset 2	5	43.45	45.35	49.80	55.44	**66.68**
	10	39.81	45.57	48.43	51.97	**55.58**
	15	33.80	34.48	39.21	39.20	**43.62**
	20	25.26	28.25	34.18	33.59	**35.76**

Table.3 Normalized Mutual Information comparision for different methods.

Dataset	N	NC	SSNLDA	NMF	AT	The proposed algorithm
Dataset 1	5	24.43	30.29	33.27	36.50	**39.21**
	10	19.72	23.15	25.85	28.96	**30.93**
	15	15.90	16.36	17.76	19.43	**23.45**
	20	12.46	16.99	18.09	18.97	**19.78**
Dataset 2	5	25.64	28.38	26.11	29.21	**33.80**
	10	17.62	19.39	20.97	25.99	**26.39**
	15	12.03	13.54	13.44	15.98	**18.98**
	20	10.03	12.43	14.53	14.85	**16.59**

Combining all the experimental results above, it can be drawn that our proposed algorithm can effectively solve the problem of detect user community for the large-scale social network than other methods. The reasons lie in that the proposed algorithm can effectively describe the relationship between users who have the same or similar interests.

V. CONCLUSIONS

This paper presents a novel large-scale social network community discovery algorithm based on probabilistic model. The user community detection problem in this paper is converted to a graph partition problem. Afterwards, Secondly, we represent the community topics as multinomial distribution on words, and user interests in different topics follows the probabilistic distribution on community topics.

REFERENCES

[1] Slusher Barbara S., Conn P. Jeffrey, Frye Stephen, Bringing together the academic drug discovery community, Nature Reviews Drug Discovery, 2013, 12(11): 811-812

[2] Berlingerio Michele, Pinelli Fabio, Calabrese Francesco, ABACUS: Frequent Pattern Mining-based Community Discovery In Multidimensional Networks, Data Mining And Knowledge Discovery, 2013, 27(3): 294-320

[3] Zhang Zhongfeng, Li Qiudan, Zeng, Daniel, User community discovery from multi-relational networks, Decision Support Systems, 2013, 54(2): 870-879

[4] Liu Jin, Zhou Jing, Wang Junfeng, Irregular community discovery for cloud service improvement, Journal of Supercomputing, 2012, 61(2): 317-336

[5] Lin Yu-Ru, Sun Jimeng, Sundaram Hari, Community Discovery via Metagraph Factorization, ACM Transactions on Knowledge Discovery from Data, 2011, 5(3), Article No.17

[6] Wang Fei, Li Tao, Wang Xin, Community discovery using nonnegative matrix factorization, Data Mining And Knowledge Discovery, 2011, 22(3): 493-521

[7] Kim Kangil, Shan Yin, Xuan Hoai Nguyen, Probabilistic model building in genetic programming: a critical review, Genetic Programming And Evolvable Machines, 2014, 15(2): 115-167

[8] Kim Younghoon, Shim Kyuseok, TWILITE: A recommendation system for Twitter using a probabilistic model based on latent Dirichlet allocation, Information Systems, 2014, 42: 59-77

[9] Gambelli, Danilo; Solfanelli, Francesco; Zanoli, Raffaele, Feasibility of risk-based inspections in organic farming: results from a probabilistic model, Agricultural Economics, 2014, 45(3): 267-277

[10] Dahal Ranjan Kumar, Bhandary Netra Prakash, Hasegawa Shuichi, Topo-stress based probabilistic model for shallow landslide susceptibility zonation in the Nepal Himalaya, Environmental Earth Sciences, 2014, 71(9): 3879-3892

[11] He Chu, Zhuo Tong, Ou Dan, Nonlinear Compressed Sensing-Based LDA Topic Model for Polarimetric SAR Image Classification, IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing, 2014, 7(3): 972-982

[12] Shen Li, Tang Hong, Chen Yunhao, A Semisupervised Latent Dirichlet Allocation Model for Object-Based Classification of VHR Panchromatic Satellite Images, IEEE Geoscience and Remote Sensing Letters, 2014, 11(4): 863-867

[13] Noel George E., Peterson Gilbert L., Applicability of Latent Dirichlet Allocation to multi-disk search, Digital Investigation, 2014, 11(1): 43-56

[14] usumaningrum Retno, Wei Hong, Manurung Ruli, Integrated visual vocabulary in latent Dirichlet allocation-based scene classification for IKONOS image, Journal of Applied Remote Sensing, 2014, 8, Article No. 083690

[15] Aubert A. H., Tavenard R., Emonet R., Clustering flood events from water quality time series using Latent Dirichlet Allocation model, Water Resources Research, 2013, 49(12): 8187-8199

[16] Zhao Bei, Zhong Yanfei, Zhang Liangpei, Scene classification via latent Dirichlet allocation using a hybrid generative/discriminative strategy for high spatial resolution remote sensing imagery, Remote Sensing Letters, 2013, 4(12): 1204-1213

[17] Zhuang Liansheng, Gao Haoyuan, Luo Jiebo, Regularized Semi-Supervised Latent Dirichlet Allocation for visual concept learning, NEUROCOMPUTING, 2013, 119: 26-32

[18] Rasiwasia Nikhil, Vasconcelos Nuno, Latent Dirichlet Allocation Models for Image Classification, IEEE transactions on Pattern Analysis and Machine Intelligence, 2013, 35(11): 2665-2679

[19] Momtazi Saeedeh, Naumann Felix, Topic modeling for expert finding using latent Dirichlet allocation, Wiley Interdisciplinary Reviews-data Mining and Knowledge Discovery, 2013, 3(5): 346-353

[20] Kwak, H., Lee, C., Park, H., & Moon, S.. What is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web, 2010, pp. 591-600.

[21] P. Jing, D. Zeng, Topic-based web page recommendation using tags, in: IEEE International Conference on Intelligence and Security Informatics, 2009, pp. 269–271.

[22] Riaz Farhan, Silva Francisco Baldaque, Ribeiro Mario Dinis, Impact of Visual Features on the Segmentation of Gastroenterology Images Using Normalized Cuts, IEEE Transactions on Biomedical Engineering, 2013, 60(5): 1191-1201

[23] Zhang H., Giles, C. L., Foley, H. C., Yen, J. Probabilistic community discovery using hierarchical latent gaussian mixture model. In Proceedings of the AAAI Conference on Artificial Intelligence, 2007, pp. 663-668.

[24] S. Zhang, R. Wang, X. Zhang, Uncovering fuzzy community structure in complex networks, Physical Review E, 2007, 76 (4), Article No. 046103.

[25] M. Steyvers, P. Smyth, M. Rosen-Zvi, T. Griffiths, Probabilistic author-topic models for information discovery, In: The Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 306–315.