Social Role Clustering with Topic Model

Jie Bai[1, 2] Linjing Li[1] Daniel Zeng[1, 2, 3] Junjie Lin[1, 2]
[1] The State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, China
[2] School of Computer and Control Engineering, University of Chinese Academy of Sciences, China
[3] Department of Management of Information Systems, The University of Arizona, USA
{baijie2013, linjing.li, dajun.zeng, linjunjie2013}@ia.ac.cn

***Abstract*—In this paper, we propose a new role analyzing paradigm for social networks enlightened by topic modeling, which can be adopted as a primitive building block in various security related tasks, such as hidden community finding, important person recognizing and so on. We first present the social network under analyzing as a heterogeneous network constructed by both the users and the subjects discussed among them. We then view this network in a Bag-of-Users schema, which mimics its classical Bag-of-Words counterpart. In this schema, the subjects discussed are treated as "documents" while the users are treated as "words" which construct the "documents". Based on this novel presentation, we finally apply topic modeling technology to perform the social role clustering. Experiments on a practical security-related social network dataset prove the effectiveness of our approach.**

Keywords—social role; social network; topic model; network structure mining; hidden community;

I. INTRODUCTION

Recently, the research of social role analysis that aims at revealing the hidden roles and behavior pattern of the user groups has been attracting growing interest in the field of both computer science and social science [1]. Generally speaking, different roles in a social network usually obey different habits in the sense of what and how they care about, which provides the possibility of discovering particular social roles and thus has great potential in social situation analysis and security related events detection. However, with the unprecedented scale and the great complexity, the networks presented in social media like Twitter and Weibo have made the precise discovering of social roles a great challenge.

Basically, to perform the social role analysis, one should explore the latent structure of social networks, which is highly correlated with the works of user interest and social interaction analysis [1]. Traditional social network analysis methodologies are mostly focused on the underlying graph structure [2], which cannot make a full use of the rich heterogeneous knowledge and text content generated from the social media. Statistical topic models, on the other hand, are advanced in discovering latent structure from a large amount of complex data, and have been extensively studied in various Natural Language Processing (NLP) problems. In a typical topic model, each latent topic is defined as a proportional distribution over words, and each document can be presented as a proportional distribution over topics. In this paper, we make attempt to apply a classical topic model named Latent Dirichlet Allocation (LDA) in mining latent structure of the social network, and cluster the users according to their potential social roles.

In recent years, some studies applied topic models for better analysis of the social roles through content generated by the users. McCallum et al. proposed an Author-Recipient-Topic model [4], which models the per-message topic distribution on both the authors and the recipients jointly. They further applied the person-conditioned topic distributions to role discovering. Zhao et al. [1] incorporated Social Role Theory into a generative process of social media content, and proposed a regularized topic model for content based social role modeling. Besides, Sun et al. proposed a ranking-based clustering algorithm for heterogeneous information networks [5], which also models "topic" for different kinds of nodes.

Different from the works mentioned above, we present a novel method to present social networks in a Bag-of-User (BoU) schema which simulates the Bag-of-Words (BoW) model in text presentation, and perform social role clustering through LDA directly. Our basic assumption is that the individuals in a social network who share the same role characters usually have similar subject interest and behavior pattern reflected in their behaviors for different discussions. Thus the users from a specific social role group usually behavior similarly for most subject discussed, like the ratio they post tweets. It should be noticed that the "subject" we discuss here means the abstraction of user interest, while "topic" we mentioned above means the hierarchical output of the topic models, which is also equivalent to the "social role cluster" that will appear in the next section.

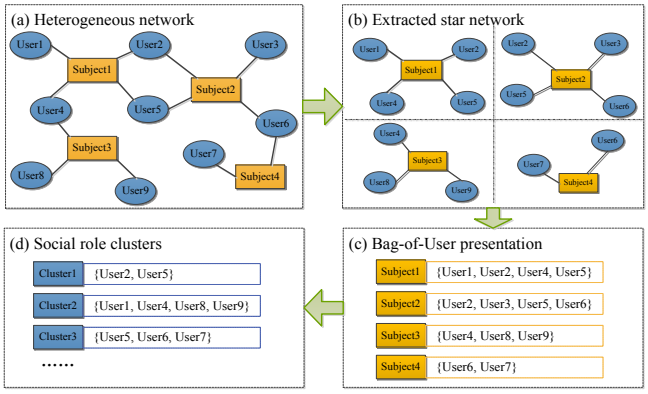
We construct a subject-user heterogeneous network for social media data, and then transform the network into a BoU presentation, where the subjects act like "documents" and the users act like "words". Based on this presentation schema, we cluster the users according to their social roles using LDA, which has been proved to be very effective in discovering the latent sematic of the words through their structural co-occurrence relationships. We conduct experiments on a massive Weibo dataset, the experimental results show that our approach clusters users according to their roles effectively.

The rest of this paper is organized as follows. In Section 2 we introduce the BoU schema and how LDA is applied in this schema. Section 3 presents the details of the experiments and Section 4 concludes the paper.

II. Social Role Clustering

In this section, we first explain the structure of the social network used in our study and the approach of presenting the social network by BoU schema. Then we demonstrate how LDA can be used in social role clustering. Figure 1 illustrates the overall workflow of our approach.

Fig. 1. Schematic diagram of the overall workflow.



A. Bag-of-User Presentation for Social Network

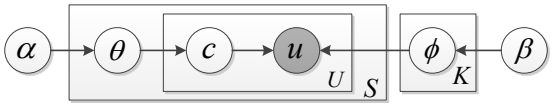
A social network here refers to the network consisting of the user and the tweets they generate in social media during a particular time period. To make things more clearly, we map the tweets into a space constructed by the subjects they involved. Then a subject-user heterogeneous network can be constructed, which is actually a bipartite graph involving users and subjects, like what illustrates in Figure 1(a). The weight of a particular edge in this network refers to the frequency that the user posts tweets belonging to corresponding subject.

BoW is a basic text presentation approach in most NLP frameworks, and it is also the input schema for LDA model. In BoW, documents are mapped into a vector space, where each dimension, corresponding to a particular word, reflects the frequency that the word appears in the document. In order to build LDA model to cluster the users, we need to transform the heterogeneous network into a BoW-like schema, that's what we call BoU. Firstly we extract the subject-centered star networks from the heterogeneous network. More specifically, we extract all of the subject-user pairs and reconstruct them according to the subjects, as is illustrated in Figure 1(b). For each subject in the network, a corresponding weighted user list are generated according to the star networks, as is illustrated in Figure 1(c). Up to now, the subject-user network has been transformed into BoU presentation.

B. LDA for Social Role Clustering

LDA is a three-layer hierarchical Bayesian generative model. It posits that each word in the documents is generated by a randomly chosen topic, and the topics are drawn from a Dirichlet distribution. Following the same idea as modeling topics for words in text, we use LDA to model the hidden social role clusters for users in the social network. In this scenario, the input of the LDA model is the social network presented in BoU schema (as is illustrated in Figure 1(c)), the output of the model are the social role clusters (as is illustrated in Figure 1(d)). The generative process of LDA applied in BoU schema can be illustrated in Figure 2, which shows that LDA learn K social role clusters from S subjects, and the user size is U. α and β are the hyper parameters which control the Dirichlet distributions.

Fig. 2. LDA applied in BoU schema



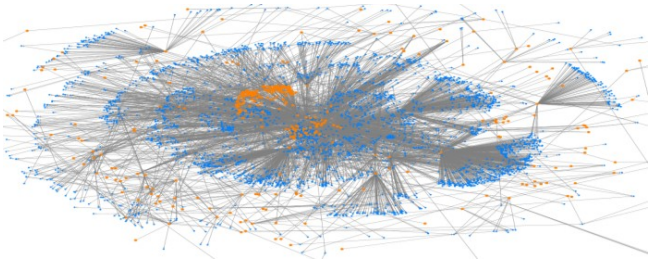
III. Experiments

In this section, we first describe the dataset used and experimental setup, then explain the experimental results in detail.

A. Data Collection and Setup

We collect massive security-related social media data from Sina Weibo, the largest Twitter-like social media platform in China. During the time from January 2011 to May 2016, we collected 14,340,675 tweets referring to sensitive subject keywords, these tweets come from 3,804,548 users. According to whether the specific keyword is included in the tweet, these tweets can be divided into 3,711 subjects. The origin tweets and retweets makes no difference here in the experiment. Considering there are also some inactive users in Sina Weibo, we filtered the users whose tweets are less than 10 in the dataset. Thus we got 7,164,537 tweets ultimately, which come from 241,930 users and belong to 3,457 subjects. The subject-user heterogeneous network for the experimental data is plotted in Figure 3 by NodeXL[1], where the orange nodes represent for the subjects, and the blue ones represent for the users.

Fig. 3. Subject-user heterogeneous network of the experimental dataset



The LDA model is trained through Gensim[2]. During the experiment, our proposed approach is denoted as "BoU-LDA", while we compare our method with the ranking-based clustering algorithm NetClus [5].

B. Results

The experiments are conducted mainly from two parts, which compare social role clustering results through both user distributions in each cluster and the user maps on the graph.

In the first part, we check the user information in each cluster directly. Following the previous study in topic models

[1] http://nodexl.codeplex.com/
[2] http://radimrehurek.com/gensim/models/ldamodel.html

and considering the practical scenes in social media, we set the topic number K = 50. Table 1 shows the user clustering results by BoU-LDA and NetClus. For a better presentation and understanding, we choose 5 representative clusters containing primarily official or business accounts, shield some sensitive accounts and list their top 5 users separately. While out of presentation, there are still some clusters consisted by personal accounts who involve similar subjects but cannot be distinguished from the username literally.

TABLE I. SOCIAL ROLE CLUSTREING RESULTS

Cluster	User Lists in Representative Clusters	
	BoU-LDA	***NetClus***
Public opinion related	金鸽舆情监测中心 (Gold Pigeon PO[a] survey) 新华网舆情在线 (Xinhua PO) 军犬舆情 (Armydog PO) 天涯舆情 (Tianya PO) 大象舆情 (Elephant PO)	-
Official news	北京青年报 (Beijing Youth Daily) ZAKER 新闻频道(ZAKER News) 新京报 (Beijing News) 财经网 (Caijing.com.cn) 四川广播电视台 (Sichuan Radio & TV Station)	头条新闻 (Headline News) 人民网 (people.cn) 财经网 (Caijing.com.cn) 北京青年报 (Beijing Youth Daily) 中国日报 (China Daily)
Market & finance	中国政府网 (The Gov Website of PRC) 环球市场播报 (Global Market Broadcast) 财经国家周刊 (Economy & National Weekly) 新浪财经 (Sina Finance) 瞭望东方周刊 (Oriental Outlook)	财经网 (Caijing.com.cn) 财新网(Caixin.com) 新浪财经 (Sina Finance) 香港商報網 (Hong Kong Commercial Daily) 瞭望 (Outlook Weekly)
IT & data analysis	股票数据分析(Stock Data Analysis) 电商叔叔 (Uncle E-commerce) 新浪科技 (Sina Science & Tech) IT 大数据挖掘 (IT & Data Mining) 电商头条 (E-commerce Headline)	-
Auto related	四川国盛奔腾 4S 店 (Guosheng Besturn[b] 4S store in Sichuan) 奔腾江北汽车城店 (Besturn store in Jiangbei Auto City) 广西钜荣奔腾 4S 店 (Jurong Besturn 4S store in Guangxi) 博瑞祥和奔腾 (Borui Xianghe Besturn store) 河北盛美奔腾 (Shengmei Besturn 4S store in Hebei)	一汽奔腾 (FAW Besturn) 奔腾江北汽车城店 (Besturn store in Jiangbei Auto City) 河北盛美奔腾 (Shengmei Besturn 4S store in Hebei) 润华一汽奔腾 (Runhua FAW Besturn store) 汽车控_小郁 (Auto fan Xiaoshao)

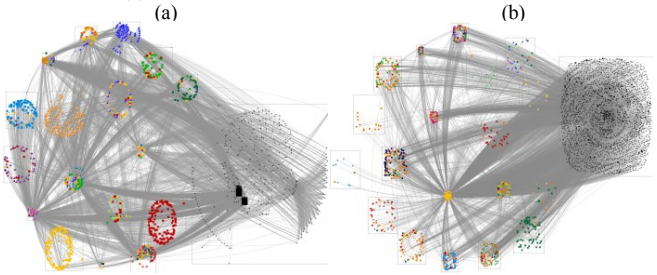
[a] "PO" here stands for "public opinion"

[b] "Besturn" is the brand of a Chinese car series

Table 1 shows that BoU-LDA effectively clusters the users according to their social role. However the clustering results of Netclus neither cover the groups mentioned in Table1, nor does it include other useful groups explicitly. The experimental results show that BoU-LDA is more appropriate for the practical scenes where noise and data imbalance may happen.

To test the diversity and novelty of the clustering results, we build a user network based on the subject-user network and map the user clusters on it. Firstly we calculate user interactions according to the times they post tweets in same subjects, then we select the top 10,000 active user interaction pairs and plot them on the graph. Based on this user network, the user clusters are plotted into groups and labeled with different colors separately. Here we set K = 20 for a clearer presentation. Figure 4 shows the user clusters mapped on the graph, in which the black nodes represents the users whom are not included in any group. In the left graph where users are clustered through BoU-LDA, there are less black nodes left and more interactions among the user groups. While in the right graph where users are clustered through NetClus, there are more black nodes and less group-wide interactions, and the graph has a "center group" which connects other user groups. It is obviously that the left graph shows more diversity and structural insights of the user network.

Fig. 4. User cluster maps, the clustering approaches are BoU-LDA (a) and NetClus (b)



IV. CONCLUSION

In this paper, we proposed a paradigm of mining latent structure of social network by topic models. We presented the subject-user heterogeneous network into BoU schema, and clustered users based on the concerned subjects and behavior patterns for different social roles. We conducted experiments on a security-related dataset, and from the encouraging first results we can see the potential of applying topic models in mining latent structure of social networks.

Improvements could be made from the following perspective. First, the information of social network utilized in this study is simple, and more relationships such as retweeting and following can be taken advantage of. Second, there is no gold standard in the experiment since we use a massive dataset collected by ourselves, which leads to the experiments focused on qualitative analysis. To solve this problem, public datasets could be used in the future study. Finally, efforts should be made to improve the existing topic models for a better adaptability of the social network structure analysis.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants 71202169, 71402177, 71602184, 71621002 and U1435221, as well as the Early Career Development Award of SKLMCCS.

REFERENCES

[1] W. X. Zhao, J Wang, Y He, J. Y. Nie, J. R. Wen, and X Li, "Incorporating social role theory into topic models for social media content analysis," IEEE Transactions on Knowledge and Data Engineering, Vol. 27, pp. 1032-1044, April 2015.

[2] D. J. Watts, "Six degrees: the science of a connected age," WW Norton & Company, 2004.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, Vol. 3, pp. 993-1022, May 2003.

[4] A. McCallum, C. Andres, and X. Wang. "Topic and role discovery in social networks," Computer Science Department Faculty Publication Series Vol. 3, 2005.

[5] Y. Sun, Y. Yu, and J. Han. "Ranking-based clustering of heterogeneous information networks with star network schema," Proceedings of the 15th ACM SIGKDD, July 2009.