

Chapter 142

Detection of Topic Communities in Social Networks Based on Tri-LDA Model

Wei Ou, Zhanfu Xie, Xiping Jia, and Binbin Xie

Abstract Social networks, in particular microblogs, have gained huge popularity in recent years. The detection of topic communities in social networks carries high value in commercial promotion, public opinion monitoring, etc. There are some existing algorithms that can detect topic communities very well. In this chapter we propose a new approach by using probabilistic generative topic model LDA (Latent Dirichlet Allocation): we add a modification to LDA to get Tri-LDA model, to process the data of friendship between users in a social network for detection of topic communities. The experiment result shows that the topic communities found by Tri-LDA are basically consistent with the realistic topic communities that are hand-labeled by the authors in the test data set.

Keywords LDA • Topic community discovery • Social networks

142.1 Introduction

The most widely used probabilistic topic model in the field of text mining is pLSI (probabilistic Latent Semantic Indexing) proposed by Hoffman [1] and LDA (Latent Dirichlet Allocation) proposed by Blei [2]. pLSI is a generative model that can be used to discover the mixture weights of topics of documents in a training data set; however, it is limited by the fact that it cannot properly infer the topic mixture weights of unseen documents. Based on the principle of pLSI, Blei proposed LDA that addressed this issue. LDA is a fully generative model that patterns each document m as a distribution $\vec{\theta}_m$ over all topics K and each topic k as a distribution $\vec{\varphi}_k$ over all words in the vocabulary V . LDA first draws $\vec{\theta}_m$ and $\vec{\varphi}_k$ from Dirichlet distributions, then samples a topic for each word position in each document m in the corpus according to $\vec{\theta}_m$, and finally emits relative words from the topic-word distribution $\vec{\varphi}_k$.

W. Ou (✉) • Z. Xie • X. Jia • B. Xie
School of Computer Science and Technology, Guangdong Polytechnic Normal University,
510665 Guangzhou, China
e-mail: studyouwei@gmail.com

LDA has been widely applied to text mining, digital image processing, etc. Usually the current applications of LDA are always “content-based” that process the content of text or pixels in images. To our best knowledge, LDA has not been used to handle social network structure data (in the context of this chapter “social network structure data” means data of friendship or relation between users in a social network; in the following sections “network structure data” and “friendship data” are used interchangeably) to detect the underlying topic communities in a social network. In this chapter, we attempt such a new application by adding a slight modification to LDA.

In the following sections, we first specify the incapability of LDA in handling social networks’ friendship data, and then we introduce our model—Tri-LDA to address it. Lastly we test the effectiveness of our method through an experiment.

142.2 Tri-LDA Model

Figure 142.1 (left) displays the Bayesian network of LDA [2]. By borrowing the generation process of documents in the LDA model specified by Blei [1] and Heinrich [3], we interpret LDA’s generation of the friendship data in a social network with N users (each user has arbitrary N_u friends) and K topic communities as follows:

- i. Sampling topic probability mixture $\vec{\theta}_u$ for each user u from the Dirichlet distribution with a prior of \vec{a} , where $\vec{\theta}_u = \{\theta_{u,k}\}_{k=1}^K$ represents the degree to which u likes about each topic k . At the meantime, sampling a distribution $\vec{\varphi}_k$ for each topic k over all the users from the Dirichlet distribution with a prior of $\vec{\beta}$, where $\vec{\varphi}_k = \{\varphi_{k,u}\}_{u=1}^U$ represents the possibility of each user being added by others as a friend in topic k .
- ii. For the n^{th} position in user u ’s friend list that has N_u positions for potential friends: firstly, sampling a topic z_a ($a = [u, n]$) for the n^{th} friend from $\vec{\theta}_u$ (this represents that user u will add a new friend who is active in this very topic for this friend position) and secondly, sampling a friend f_a according to
$$p(f_a | \vec{\varphi}_{z_a}, z_a) = \varphi_{z_a, f_a}.$$
- iii. Repeat step 2 till all friends of all users are generated.

However, it can be easily found that $\vec{\theta}_u$ and $\vec{\varphi}_k$ are correlated, instead of being independent of each other: for an arbitrary topic k , the more a user likes about it, the more likely this user could be added by other users who also like the topic k to their friend lists. Or, namely, the higher the probability $\theta_{u,k}$, the higher $\varphi_{k,u}$ is.

We solve this correlation problem by adding a new parameter that is directly proportional to $\vec{\theta}_u : \vec{\pi}_u^{(k)}$ that represents the degree to which u accept other users from

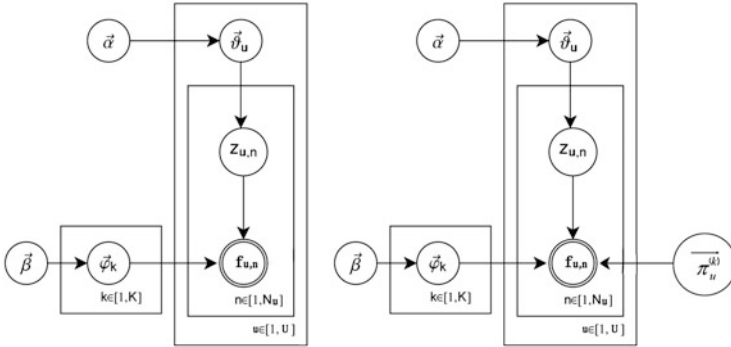


Fig. 142.1 Bayesian networks of LDA (left) and Tri-LDA (right)

the same social network in topic k as his or her friends. In the meantime, we give $\vec{\varphi}_k$ a new meaning to make the generation process more natural: the activeness of each user in topic k . The possibility of an arbitrary user u adding another user who is active in topic z_a to the n^{th} position in his or her friend list as friend f_a is determined by the joint distribution of $\vec{\varphi}_{z_a}$ and $\vec{\pi}_u^{(z_a)}$, which is expressed by the following equation:

$$p\left(f_a \mid \vec{\varphi}_{z_a}, z_a, \vec{\pi}_u^{(z_a)}\right) = \varphi_{z_a, f_a} \pi_{u, f_a}^{(z_a)} \tag{142.1}$$

By adding this slight modification, we get the modified LDA model—Tri-LDA. Figure 142.1 (right) displays the Bayesian network structure of Tri-LDA. In this chapter we simplify $\vec{\pi}_u^{(k)}$ as: $\pi_{u_1, u_2}^k = \begin{cases} 0 & \text{if } u_1 = u_2 \\ \theta_{u_2, k} & \text{if } u_1 \neq u_2 \end{cases}$

That means, firstly, no user would add himself or herself to his or her friend list and, secondly, the degree to which user u_1 accepts another user u_2 as a friend in topic k equals to $\theta_{u_2, k}$, the degree to which user u_2 likes about topic k . Therefore based on the simplification, we can get the following equation:

$$p\left(f_a \mid \vec{\varphi}_{z_a}, z_a, \vec{\pi}_u^{(z_a)}\right) = \varphi_{z_a, f_a} \theta_{f_a, z_a} \tag{142.2}$$

There are two methods that are commonly used to infer latent parameters for high-dimensional probabilistic models like LDA: EM (Expectation Maximization) and Gibbs sampling. In this chapter, for Tri-LDA, we choose Gibbs sampling to infer $\vec{\theta}$, the matrix that represents the degree to which each user likes about each topic in a social network.

Gibbs sampling is a simple implementation of the Monte Carlo algorithm, which is usually used to infer the latent parameters of high-dimensional probabilistic topic

models [4]. Gibbs sampling starts from a randomly sampled initial state, then continuously transits to new states based on the current state to build up a Markov chain. After a certain number of transitions, the Markov chain would become stable. The stable state of the Markov chain can be regarded as an approximated observation of the query distribution. In this chapter we run Gibbs sampling methods on the social network to get an approximated observation of $\vec{\theta}$: the topic indices \vec{z} from which users add their friends to his or her friend list. We specify the implementation of Gibbs sampling on a social network as the following steps:

- i. Randomly allocate a topic k for every friend in every user's friend list.
- ii. In an arbitrary user u 's friend list, choose the n^{th} friend position, indexing it with i . Remove the topic k that previously allocated on f_i , and then resample a new topic for it based on the following two known conditions: all the friends of all users \vec{f} and all the topics of all friends in all users' friend lists except for the topic of f_i : z_i .
- iii. Repeat step ii through all friends in user u 's friend list.
- iv. Repeat step iii through all users in the social network.

In step 2, the posterior probability $p\left(\vec{z}_i | \vec{z}_{-i}, \vec{f}\right)$ needs to be computed. According to Tri-LDA stated above, we can get the following equation:

$$p\left(\vec{f}, \vec{z} | \vec{\theta}, \vec{\varphi}, \vec{\pi}\right) = \prod_{u=1}^U \prod_{k=1}^K \theta_{u,k}^{n_u^{(k)} + n_{u'}^{(k)}} \prod_{k=1}^K \prod_{u=1}^U \varphi_{k,u}^{n_{u'}^{(k)}} \tag{142.3}$$

where n_u^k denotes the number of friends who are allocated with topic k in the friend list of user u and $n_{u'}^k$ denotes in the friend lists of other users the number of times that user u is allocated with topic k . Then, we get the following equation:

$$p\left(\vec{f}, \vec{z} | \vec{a}, \vec{\beta}, \vec{\pi}\right) = \prod_{k=1}^K \frac{\Delta\left(\vec{n}_{u'} + \vec{\beta}\right)}{\Delta\left(\vec{\beta}\right)} \prod_{u=1}^U \frac{\Delta\left(\vec{n}_u + \vec{a} + \vec{n}_{u'}\right)}{\Delta\left(\vec{a}\right)} \tag{142.4}$$

where $\vec{n}_{u'} = \left\{n_{u'}^{(k)}\right\}_{k=1}^K$, $\vec{n}_u = \left\{n_u^{(k)}\right\}_{k=1}^K$. Assume the i^{th} friend in u 's friend list is user u_o , and then we get the following posterior distribution $p\left(\vec{z}_i | \vec{z}_{-i}, \vec{f}\right)$:

$$\begin{aligned}
p\left(z_i = k \mid \vec{z}_{-i}, \vec{f}\right) &\propto \frac{p\left(\vec{f} \mid \vec{z}\right) p\left(\vec{z}\right)}{p\left(\vec{f}_{-i} \mid \vec{z}_{-i}\right) p\left(\vec{z}_{-i}\right)} \\
&\propto \frac{n_{u_o, -i}^{(k)} + \beta_{u_o}}{U} * \left(a_k + n_{u, -i}^{(k)} + n_{u'}^{(k)}\right) * \left(a_k + n_{u_o}^{(k)} + n_{u_o, -i}^{(k)}\right) \\
&\quad \sum_{u=1} n_{u', -i}^{(k)} + \beta_u
\end{aligned} \tag{142.5}$$

where $n_{u_o, -i}^{(k)}$ denotes the number of times that user u_o is allocated with topic k excluding i in all other users' friend lists and $n_{u, -i}^{(k)}$ denotes the number of friends who are allocated with topic k in the friend list of user u excluding i . After running Gibbs sampling on the social network, we get all the topic indices from which all friends of all users are sampled that can be regarded as an approximated observation of $\vec{\theta}$. With the approximated observation, we can compute the expectation of the Dirichlet distribution and use it as an estimator for the desirable $\theta_{u,k}$, the degree to which a user u likes about topic k (Table 142.1).

$$\theta_{u,k} = \frac{n_u^{(k)} + a_k + n_{u'}^{(k)}}{\sum_{k=1}^K n_u^{(k)} + a_k + n_{u'}^{(k)}} \tag{142.6}$$

142.3 Experiment

We collect a data set that includes the connections between 2,315 Sina Weibo users. The friends of each user are limited to be the users in this data set. By reviewing the data set, we find that around 70% of nodes (or users) are richly connected with each other, and around 10% of nodes are relatively isolated. We randomly choose 2,179 of them as the training data set and the remaining as test data set. By manually checking the posts and tags those users posted on the site, we find there are 17 different topics involved in the data set: information technology, business, finance, military, charity, food, education, car, travel, photography, everyday life information, show business, politics, sports, literature, painting, and religion. Based on the contents of each user's post page, we score the degrees to which each user is interested in those topics. By using the scores, we represent each user's interest by a vector with 17 elements: \vec{h}_u . Then the similarity of interests between two users u_1, u_2 can be expressed as:

$$\text{Sim}_{u_1, u_2} = \frac{|\vec{h}_{u_1} \cdot \vec{h}_{u_2}|}{|\vec{h}_{u_1}| |\vec{h}_{u_2}|} \quad (142.7)$$

In this experiment, we first run Gibbs sampling based on the Tri-LDA model to discover the underlying topic communities. We measure the effectiveness of Tri-LDA by computing the fittingness of the outputted result to the actual topic communities manually labeled by us. We also run k -medoid clustering algorithm on the texts and tags posted by the users and see users of each cluster in the output as a topic community [5]. To test the generalization property of Tri-LDA, we compare the predictive perplexity of Tri-LDA with that of k -medoid clustering mentioned above.

In the training process of LDA, we set the prior parameter of a in the range [0.1, 1.5] with an incremental step of 0.1, the prior parameter of β in the range [1.0, 10.0] with an incremental step of 0.1, and the topic number K in the range [5, 20] with an incremental step of 1. We run Gibbs sampling with all the possible parameter combinations and select the one whose output has the best fittingness value. After the learning process, the desirable topic probability mixture $\theta_{u,k}$ for every user u , which represented the degree to which user u likes about a topic k , is outputted. We set a threshold value 0.25 for $\theta_{u,k}$: if $\theta_{u,k} \geq 0.25$, then we conclude that u is interested in topic k . Denoting T_k as the set of users in the k^{th} topic community, then $T_k = \{u | \theta_{u,k} \geq 0.25\}$.

In the real-world social networks, users from the same topic community always have relatively high interest similarities. To measure the credibility of the learning result, we use the interest similarity equation stated above to compute the interest similarity between each two users in a topic community obtained by Tri-LDA and take its averaged value to measure the fittingness of the learning result to the real-world situation. We express the fittingness of the learning result as the following equation:

Table 142.1 Pseudocode of Gibbs sampling algorithm for Tri-LDA

Initialization

for all users $u \in [1, U]$ in the social network do
 for all friend positions $n \in [1, N_u]$ in user u 's friend list do
 randomly allocate n with a topic $k \in [1, K]$

Building a Markov Chain

while not convergent do
 for all users $u \in [1, U]$ in the social network do
 for all friend positions in user u 's friend list do
 resample n with a new topic k according to equation (142.5)

Parameter estimation

for every user u in the social network do
 compute $\vec{\theta}_u$ according to equation (142.6)

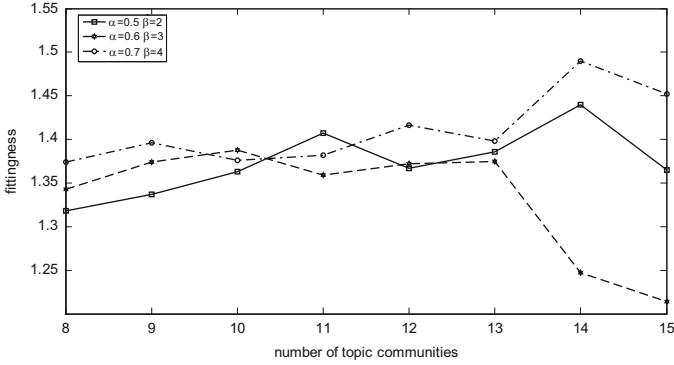


Fig. 142.2 Fittingness of the Tri-LDA model

$$f = \frac{\sum_{k=1}^K \text{avg}(\text{Sim}_k)}{K} \tag{142.8}$$

where $\text{avg}(\text{Sim}_k)$ denotes the average interest similarity between every two users in a topic community. In the experiment we find that when $a = 0.7, \beta = 8, k = 14$, the topic communities outputted by the algorithm have the best fittingness value. Figure 142.2 displays the fittingness value of the learning results under some of the parameter settings in this experiment. The result shows that the average fittingness f under different parameter settings in this experiment is around 0.70, which indicates the learning results can basically reflect the actual interest similarities between users in the selected social network.

We denote the degree to which a user d in the test data set likes about topic k as $\theta_{d,k}$, then

$$\theta_{d,k} = \frac{n_d^{(k)}}{N_d} \tag{142.9}$$

where $n_d^{(k)}$ is the number of friends who belong to topic community k in d 's friend list and N_d is the total number of friends in d 's friend list. Use the learning result to predict the interest of users in the test set, and express the predictive perplexity of the predication with the following equation:

$$\text{perplexity} = \frac{\sum_{d=1}^D H(d)}{D} \tag{142.10}$$

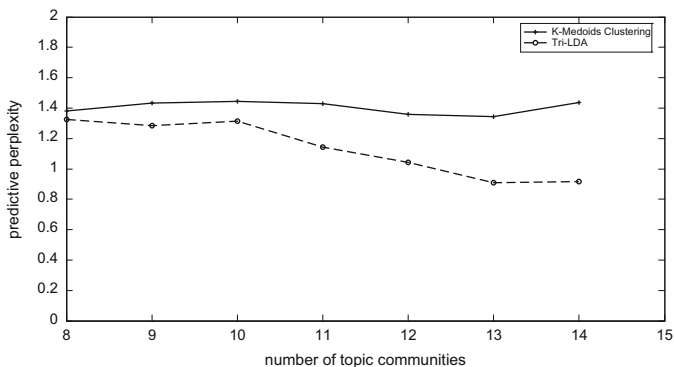


Fig. 142.3 Comparison of predictive perplexity between Tri-LDA and k -medoid clustering

where $H(d) = -\sum_{k=1}^K \theta_{d,k} \log \theta_{d,k}$ and D is the total number of users in the test data

set. A relatively lower predictive perplexity always suggests better generalization property. We compare Tri-LDA's predictive perplexity, when $a=0.7$, $\beta=8$, $k=8, 9, 10, 11, 12, 13, 14$, with the predictive perplexity of k -medoid clustering with the medoid number as 8, 9, 10, 11, 12, 13, 14, respectively. Figure 142.3 displays the comparison between the predictive perplexity of Tri-LDA and k -medoid clustering in this experiment. The result indicates that Tri-LDAs have a lower predictive perplexity than k -medoid clustering; therefore Tri-LDA has a better generalization performance that can be used to predict the interests of unseen users than k -medoid clustering in this experiment.

Conclusion and Future Work

We propose a modified LDA model—Tri-LDA—to detect topic communities in social networks by processing network structure data. The experiment result shows that the learning result is consistent with the realistic topic communities hand-labeled by us in the test data set. Also the experiment result shows that Tri-LDA has a decent generalization performance in predicting the interests of unknown users. In the future we plan to add some further modification to Tri-LDA to allow it to process the combined data of network structure and communications between users to detect the underlying topic communities in a social network.

References

1. Hofmann T. Probabilistic latent semantic analysis. In: Proceedings of the 15th conference on uncertainty in artificial intelligence. Stockholm: Morgan Kaufmann; 1999. p. 289–96.
2. Blei DM, Ng A. Latent Dirichlet allocation. *J Mach Learn Res.* 2003;3:993–1022.
3. Heinrich G. Parameter estimation for text analysis[DB/OL]. 2005. faculty.cs.byu.edu
4. LianWen Z. Introduction to Bayesian networks [M]. Beijing: Science Press; 2006 (In Chinese).
5. Li X. Tag-based social interest discovery. In: Proceeding WWW '08 Proceedings of the 17th international conference on World Wide Web. US: ACM; 2008. p. 675–84.