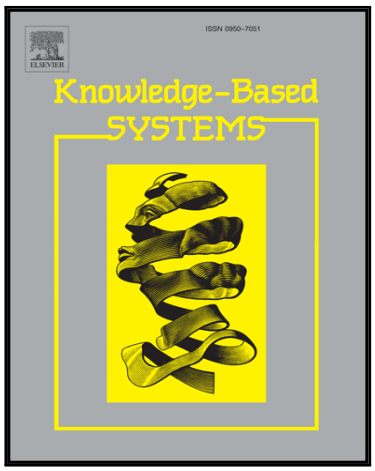

Forum Latent Dirichlet Allocation for User Interest Discovery

Chaotao Chen, Jiangtao Ren

PII: S0950-7051(17)30172-7
DOI: 10.1016/j.knosys.2017.04.006
Reference: KNOSYS 3886

To appear in: *Knowledge-Based Systems*

Received date: 27 November 2016
Revised date: 12 April 2017
Accepted date: 13 April 2017

Please cite this article as: Chaotao Chen, Jiangtao Ren, Forum Latent Dirichlet Allocation for User Interest Discovery, *Knowledge-Based Systems* (2017), doi: 10.1016/j.knosys.2017.04.006

Forum Latent Dirichlet Allocation for User Interest Discovery

Chaotao Chen, Jiangtao Ren*

School of Data and Computer Science, Sun Yat-sen University, Guangzhou, Guangdong, P.R.China 510006

Abstract

The popularity of online forums provides a good opportunity to learn user interests which can be used in many business scenarios, such as product or news recommendation. There exist many approaches to infer forum topics and users' interests. Among them, Author-Topic (AT) like models are most popular. But a thread in online forum is composed of a root post and some response posts which may be relevant or irrelevant to the root post. So the assumption of AT that response posts are generated from user's interest topics is not comprehensive. In this paper, we distinguish user's serious and unserious interest topics and argue that the topic of a relevant response post is jointly determined by its author's serious interest topics and the topics of its root post, while the topic of irrelevant response post is only determined by its author's unserious interest topics. Based on these assumptions, we propose Forum-LDA to model the generative process of root post, relevant and irrelevant response posts jointly. Therefore, our model can not only learn more coherent topics and serious interests, but also identify unserious users who publish many irrelevant posts. Extensive experiments on real forum dataset demonstrate the advantages of our model in tasks such as user interest and unserious user discovery.

Keywords: User Interest, Topic Model, Forum Content Analysis

*Corresponding author

Email addresses: chencht3@gmail.com (Chaotao Chen), issrjt@mail.sysu.edu.cn (Jiangtao Ren)

1. Introduction

With the prevalence of online social media, such as online forums, Facebook, Twitter and Question-Answering Communities, people are more and more accustomed to express and share their opinions on the Internet. As one of the most centralized platforms for online communications, online forum is receiving more and more attentions from research area. In online forum analysis, the user interests are greatly concerned, because accurate modeling of user interests is not only helpful to supplement forum content analysis [1], but also beneficial to many important applications, including user profiling [2, 3, 4], community discovery [4] and collaborative recommendation [5, 6].

Latent topic modeling approaches such as probabilistic Latent Semantic Indexing (pLSI) [7] and Latent Dirichlet Allocation (LDA) [8], are widely used in social media analysis due to their merits of simplicity, robustness and interpretability. Similarly, many related works adopt and extend the topic models to explore user interests from forum contents. For instance, Author-Topic model aggregates a user's posts and assumes a user-level distribution over topics to capture user's topical interests [2].

However, online forum is more complex than other social media, so the assumption of Author-Topic like models that the topics of a user's post is only determined by the user's interest topic distribution is not comprehensive. For instance, while some user seriously state their opinions, there are also many people playing jokes and publishing meaningless response posts that do not relate to the discussed topics. Some examples of serious/unserious interests and posts are illustrated in Table 1. So in Author-Topic like models, these users' real topical interests may be buried by the frequent meaningless topics due to the lack of explicit separation of relevant posts and irrelevant posts. And there is also a problem that the response posts in a thread are usually short so traditional topic models may suffer from the problem of sparsity and can not model user interests accurately. In fact, a web forum is an online portal for open threads, where in each thread a user proposes a root post indicating an issue

Table 1: Examples of serious/unserious interests and posts.

Unserious interest		
Joke	Praise	Obscenity
1. Hahaha! 2. Funny, haha.	1. Good post! 2. Mark.	1. No shit! 2. Drop dead.
Serious interest		
Basketball	Football	Car
1. I think Jordan is better than Kobe.	1. China team never "disappoints" us.	1. I recommend BYD.

and others express their opinions on the issue by submitting response posts. So there is a close relationship between root post and its response posts in each thread, which we consider to be helpful in understanding the forum topics and extracting user interests.

In this paper, we propose a novel topic model named Forum-LDA to extract user's real interest topics in online forums by modeling the generative process of root posts, relevant and irrelevant response posts jointly. In Forum-LDA, we divide a user's interests into two parts, one is serious interest topics the forum usually discusses, the other is unserious interest topics that account for the irrelevant response posts. With explicitly distinguishing serious and unserious interest topics, Forum-LDA is able to identify unserious posts and users so that it can get rid of the disturbance from irrelevant response posts and discover user's real interest topics from relevant response posts. In particular, we introduce a Bernoulli distribution indicating how much a user prefers to publish irrelevant response posts and a latent variable to indicate whether a response post is irrelevant to the serious topics under discussion. For an irrelevant response post, its topic is generated from its author's unserious interest topics. For a relevant response post, we hypothesize that its topic is jointly determined by its author's serious interest topics and the topics its root post concerns. This assumption is intuitive since users usually publish serious response posts according to the

content of root posts and their own interests. By incorporating relationship between response posts and their root posts, Forum-LDA is expected to overcome the sparsity problem of response posts and learn more semantic and coherent topics. Given the observation that a response post is usually short and related to only one topic, we further assume the topic number of a response post is one. This constraint is consistent with the shortness nature of response posts and improves the compactness of the derived topics.

Contributions. The main contributions of our work are summarized as follows:

1. We propose a novel topic model for forums user interest profiling, which explicitly distinguishes user's serious interest topics that we want to discover, and unserious interest topics that account for the common irrelevant response posts. So it can detect unserious post/users and capture user's real interests more accurately;
2. Our proposed model incorporates the relationship between response posts and the corresponding root posts to overcome the sparsity problem of response posts, which results in more coherent forum topics and user interests;
3. Extensive experiments conducted on real-world forum data demonstrate that our proposed model outperforms the existing approaches in user interests profiling and unserious user/post discovery in online forums.

2. Related Work

This section provides a brief overview of related works that extend the topic modeling approach for modeling online forums with user interests profiling.

Within the text modeling area, Latent Dirichlet Allocation (LDA) [8] is still by far the most widely used topic model which represents a document as a mixture of latent topics to be inferred, where a topic is a multinomial distribution of words. Many extensions of LDA have been designed for different document collections, such as academic abstracts [9], online reviews [10, 11] and

email summaries [12]. Recently, its superiority has also been demonstrated in modeling online social media, such as microblog streams [13, 14, 15], Twitter posts [16, 17, 18] and online forums [19, 20, 21].

For user interest profiling based on topic models, the Author-Topic (AT) model [2] is one of the early attempts with each user interest profile represented as a distribution of topics, where a topic is a multinomial distribution of words as the same as LDA. Linked Topic Interest model [3] explores discussion topics and user interests in a linked manner and provides an interpretation of an interest in terms of weighted topics. However, most of the Author-Topic like models we mentioned above rely on the assumption that each post is generated from its user's interest topic distribution, which ignores the relationships between response post and the root post in the same thread. To accurately model user motivation and interests in online forums, we modify the existing Author-Topic models by incorporating the relationships between root posts and response posts, where each response post is generated from the product of its author's interest distribution and its corresponding root post's topic distribution. Therefore, different response posts are linked by their joint root post and each response post can take advantages of the richer information in their corresponding root posts.

Since most response posts are usually short, our model is also related to topic modeling of short texts in social media. In recent years many efforts have been made to overcome the sparsity of short texts. Some previous studies have tried to enrich the short texts with external resources to get more features. Twitter-Network topic model [22] makes use of the accompanying hashtags, authors, and followers network to model tweets better. Wiki-LDA [23] combines LDA, text extraction APIs and Wikipedia categories to effectively mine user's interests in Twitter. However, appropriate external resources is not always available. Thus, in recent years more efforts have been put on designing customized topic models for short texts. Biterm topic model (BTM) [24] is among the earliest works, which directly models word pairs (i.e. biterms) extracted from short texts. According to the observation that a single tweet is usually related to

a single topic, Twitter-LDA [17] assumes that each tweet is associated with only one topic based on the user's topic distribution, and each word in a tweet is generated either from the chosen topic or a background topic shared by all tweets. Inspired by these prior works, we further assume that each response post has only one topic.

Our proposed model also consider the existence of irrelevant response posts and unserious users, so our work is also related to research of spam/spammer detection in social media. A similar work in online forums is Topic-Style model [1] which introduces a new type of word distributions representing writing styles and a user-level distribution over writing styles in order to identify serious versus unserious users. Differently, our proposed model allows each user to be associated with two multinomial distributions over serious interest topics and unserious interest topics, respectively. Moreover, a Bernoulli distribution is introduced to indicate how much a user likes publishing unserious posts. Therefore, our proposed model can not only identify unserious post/user, but also capture user's serious interests and unserious interests. Another similar work is Twitter-User model [18] which uses a latent variable to indicate whether a tweet is related to its author's interest. But the latent variable is governed by a global Bernoulli distribution rather than a user-specific Bernoulli distribution.

3. Model and Inference

In this section, we present a novel topic model for forum user interest discovery. We first define the variables and notations, then we formally present our proposed Forum-LDA model. Finally, we describe the parameter learning method based on Gibbs sampling.

3.1. Preliminaries

Similar to LDA, we formally define the following terms. An online forum is a collection of threads $\mathbb{D} = \{(\boldsymbol{w_1}, \boldsymbol{R_1}), \ldots, (\boldsymbol{w_D}, \boldsymbol{R_D})\}$, each thread contains a root post that is group of words $\boldsymbol{w_d}$ from a vocabulary $\mathbb{V}$ of size V, and

the corresponding response post collection $\boldsymbol{R_d}$. The response post collection of a thread is a collection of response posts $\boldsymbol{R} = \{(\boldsymbol{r_1}, u_1), \ldots, (\boldsymbol{r_M}, u_M)\}$, each contains a group of words $\boldsymbol{r_m}$ from the same vocabulary $\mathbb{V}$, and its corresponding author u_m from a set of users $\mathbb{U}$ of size U. Next we propose our Forum-LDA to infer forum topics and user interests jointly based on the forum data.

3.2. Forum-LDA Model

Author-Topic like topic models have been widely employed to discover user interests in many online social media. Most previous works assume a document is generated from its author's interest distributions, which is not comprehensive in case of online forums. In popular online forums with millions of active users, in addition to response posts which contribute positively to a discussion by offering relevant and meaningful opinions, there are also many response posts which appear irrelevant, disrespectful or meaningless, such as spams, profanities and advertisements. These response posts are uninformative and even harmful to the user interests modeling. However, most previous efforts overlook the common existence of forum users who sometimes seriously state their opinions, and sometimes publish irrelevant posts that are not related to the topics under discussion. Therefore, without explicit identification of irrelevant response post, Author-Topic like models can not discover the user's serious interest topics accurately which is mixed with some irrelevant topics. To solve this problem, we divide a user's interest into two parts, one is serious interest topics, the other is unserious interest topics. The preference to publish irrelevant response posts can be regarded as an unserious topic of interests, compared with the serious topics the forums formally discuss. Furthermore, for each user we introduce a Bernoulli distribution to measure how much he/she prefers to publish irrelevant response posts, and a latent variable to indicate whether a response post is irrelevant to the serious topics under discussion or not. With explicit separation of serious and unserious interest topics, Forum-LDA is able to identify unserious posts versus serious posts, and focuses on relevant response posts to capture users' real interest topics.

In addition to irrelevant response posts, the shortness and sparsity nature of response post may also worsen the performance of traditional topic models. To address this problem, we propose to incorporate the relationship between response posts and their root posts in the same threads. The generation mechanism of online forum threads is that one user starts a root post covering one or several topics, and other interested users publish response posts to express their opinions on one of the topics. Therefore, the response posts in the same thread are not independent, but linked by the root post. For an irrelevant response post, its topic has little to do with the root post, so its topic is assumed to be generated from its author's unserious interest topic distribution. But for a relevant response post, it is intuitive to assume its topic is determined by its author's serious interest topics and the root post topics jointly. So we extend existing Author-Topic like models with two reasonable assumptions: (1) a response post contains only one topic due to their shortness; (2) the topic of a relevant response post is generated from the product of its user's interest distribution and its root post's topic distribution. With rich information from root posts and relationships between root post and response posts, our proposed model is expected to address the sparsity problem of response posts, and it can discover more coherent topics for correct user interest discovery.

Now we formally describe the Forum-LDA model. The graphical model is illustrated in Figure 1 and the notations we use in the model is summarized in Table 2. We assume that there are T serious topics and S unserious topics, where ϕ_t and ψ_s are the word distribution for serious topic t and unserious topic s, respectively. There are D threads, where each thread d consists of a root post $\boldsymbol{w_d}$ and M_d response posts. Each root post $\boldsymbol{w_d}$ has a topic distribution θ_d while each of its response post contains only one topic. There are U users for all response posts, where each user u has a serious topic distribution π_u, an unserious topic distribution η_u and a Bernoulli distribution parameter λ_u. Root post-topic distributions θ_d, user-topic distributions η_u and π_u are all drawn from a symmetric Dirichlet distribution with prior α for C. Topic-word distributions ϕ_t and ψ_s are both drawn from a symmetric Dirichlet distribution with prior

β. And switching variable distribution parameter λ_u is drawn from a Beta distribution parameterized by γ_0 and γ_1.

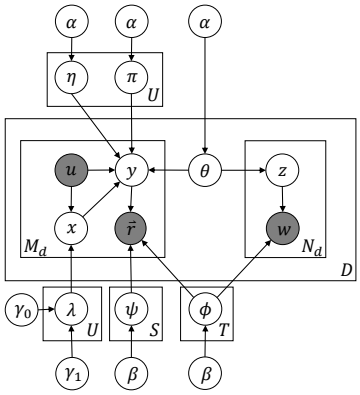

Figure 1: Graphical model of Forum-LDA.

For each word in a root post, we draw a topic z from the root post's topic distribution, and then draw the word from the corresponding serious topic-word distribution. For each response post, first a binary switching variable x is sampled from the corresponding user's Bernoulli distribution with parameter λ_u. If $x = 0$, we draw an unserious topic from the user's unserious topic distribution parameterized by η_u, and then draw the words in the response post from the corresponding unserious topic-word distribution. Otherwise, if $x = 1$, we first draw a topic from a multinomial distribution, whose parameter is the normalized product of the root post's topic distribution and user's serious topic distribution, and then draw the words in the response post from the corresponding serious topic-word distribution. The generative process of our model is described as follows:

1. For each serious topic $t = 1, 2, ..., T$, draw a multinomial topic-word distribution $\phi_t \sim Dirichlet(\beta)$.
2. For each unserious topic $s = 1, 2, ..., S$, draw a multinomial topic-word distribution $\psi_s \sim Dirichlet(\beta)$.
3. For each user $u = 1, 2, ..., U$,

Table 2: Notations

Notation	Description
α, β	Hyperparameters of Dirichlet distributions
γ_0, γ_1	Hyperparameters of Beta distribution
λ_u	User-specific Bernoulli distribution parameter for switching variable x
θ_d	Root post-specific topic distributions
π_u, η_u	User-specific serious interest topic distributions and unserious interest topic distributions
ϕ_t, ψ_s	Word distributions of serious interest topics and unserious interest topics
x, y, z	Hidden variables: x for switching, y for topic of response posts, z for topic of words in root posts
$w, \vec{r}, u$	Root post word, response post, user of response post
D, N_d, M_d	Number of threads (root posts), number of words in each root post, number of response posts in each thread
T, S, U	Number of serious interest topics, unserious interest topics and users

(a) draw a multinomial serious user-topic distribution $\pi_u \sim Dirichlet(\alpha)$.

(b) draw a multinomial unserious user-topic distribution $\eta_u \sim Dirichlet(\alpha)$.

(c) draw a Bernoulli switching variable distribution $\lambda_u \sim Beta(\gamma_0, \gamma_1)$.

4. For each thread $d = 1, 2, ..., D$,

(a) draw a multinomial root post-topic distribution $\theta_d \sim Dirichlet(\alpha)$.

(b) for each word $n = 1, 2, ..., N_d$ in the root post, draw a topic $z_{d,n} \sim Multinomial(\theta_d)$, and then draw a word $w_{d,n} \sim Multinomial(\phi_{z_{d,n}})$.

(c) for each response post $m = 1, 2, ..., M_d$ in the thread, where $u_{d,m} \in \{1, 2, ..., U\}$ is the user who has written the response post,

i. draw $x_{d,m} \sim Bernoulli(\lambda_{u_{d,m}})$.

ii. if $x = 0$, draw a topic $y_{d,m} \sim Multinomial(\eta_{u_{d,m}})$ and for each word $j = 1, 2, ..., J_{d,m}$ in the response post, draw a word $r_{d,m,j} \sim Multinomial(\psi_{y_{d,m}})$.

iii. if $x = 1$, draw a topic $y_{d,m}$ from a multinomial distribution, whose parameter is the normalized product of $Multinomial(\theta_d)$ and $Multinomial(\pi_{u_{d,m}})$, and for each word $j = 1, 2, ..., J_{d,m}$ in the response post, draw the word $r_{d,m,j} \sim Multinomial(\phi_{y_{d,m}})$.

3.3. Parameters Estimation

Exact inference to LDA type models is intractable, and Gibbs sampling [9] or variation methods are approximate algorithms for this problem. In this paper, we use Gibbs sampling to estimate the parameters. The sampling probability that assigns the nth word in the root post of thread d to serious topic t is as follow:

$$P(z_{d,n} = t | \boldsymbol{z_{\neg d,n}}, \boldsymbol{w}) \propto \frac{N_d^t + C_d^t + \alpha}{N_d + C_d + T\alpha} \cdot \frac{N_t^{w_{d,n}} + \beta}{N_t + V\beta} \quad (1)$$

where $\boldsymbol{z_{\neg d,n}}$ denotes the serious topic assignments for all words except word $w_{d,n}$, N_d^t is the number of words assigned to serious topic t in the root post of thread d, N_d is the total number of words in the root post of thread d, C_d^t is the number of response posts assigned to topic t in thread d, C_d is the total number of response posts in thread d, $N_t^{w_{d,n}}$ is the number of times that word $w_{d,n}$ is

assigned to serious topic t, N_t is the total number of times that any word is assigned to serious topic t.

The probability to assign the mth response post authored by user u in thread d to unserious topic s or serious topic t is as follows:

$$\begin{aligned} & P(x_{d,m}=0, y_{d,m}=s|\boldsymbol{x}_{\neg \boldsymbol{d},\boldsymbol{m}}, \boldsymbol{y}_{\neg \boldsymbol{d},\boldsymbol{m}}, \boldsymbol{w}, u) \\ \propto \quad & (R_u^0+\gamma_0)\cdot\frac{A_{u,0}^s+\alpha}{A_{u,0}+S\alpha} \\ & \frac{\Gamma(N_s+V\beta)}{\Gamma(N_s+C_s+V\beta)}\prod_{v=1}^{V}\frac{\Gamma(N_s^v+C_s^v+\beta)}{\Gamma(N_s^v+\beta)} \end{aligned} \tag{2}$$

$$\begin{aligned} & P(x_{d,m}=1, y_{d,m}=t|\boldsymbol{x}_{\neg \boldsymbol{d},\boldsymbol{m}}, \boldsymbol{y}_{\neg \boldsymbol{d},\boldsymbol{m}}, \boldsymbol{w}, u) \\ \propto \quad & (R_u^1+\gamma_1)\cdot\frac{A_{u,1}^t+\alpha}{A_{u,1}+T\alpha}\cdot\frac{N_d^t+C_d^t+\alpha}{N_d+C_d+T\alpha} \\ & \frac{\Gamma(N_t+V\beta)}{\Gamma(N_t+C_t+V\beta)}\prod_{v=1}^{V}\frac{\Gamma(N_t^v+C_t^v+\beta)}{\Gamma(N_t^v+\beta)} \end{aligned} \tag{3}$$

where $\boldsymbol{x}_{\neg \boldsymbol{d},\boldsymbol{m}}$ and $\boldsymbol{y}_{\neg \boldsymbol{d},\boldsymbol{m}}$ denote the indicator variable assignments and topic assignments for all response posts except response post $r_{d,m}$, respectively; R_u^0 and R_u^1 are the numbers of user u's response post assigned to unserious topic and serious topic, respectively; N_s^v and N_t^v are the numbers of times that word v is assigned to unserious topic s and serious topic t, respectively; N_s is the total number of times that any word is assigned to unserious topic s; $A_{u,0}^s$ and $A_{u,1}^t$ are the numbers of words assigned to user u's unserious topic s and serious topic t, respectively; $A_{u,0}$ and $A_{u,1}$ are the total numbers of words assigned to user u's unserious topics and serious topics, respectively; C_s^v and C_t^v are the numbers of times that word v is assigned to unserious topic s and serious topic t in response post $r_{d,m}$, respectively; C_s and C_t are the numbers of times that any word is assigned to unserious topic s and serious topic t in response post $r_{d,m}$, respectively; Γ is the Gamma function.

After Gibbs sampling, with the counters of the sampled topic assignments and indicator variable assignments, we can estimate the parameters. They can be calculated by the equations below:

$$\psi_s^v=\frac{N_s^v+\beta}{N_s+V\beta} \qquad \phi_t^v=\frac{N_t^v+\beta}{N_t+V\beta} \tag{4}$$

$$\theta_d^t = \frac{N_d^t + C_d^t + \alpha}{N_d + C_d + T\alpha} \qquad \lambda_u^x = \frac{R_u^x + \gamma_x}{\sum_{x'=1}^{1} R_u^{x'} + \gamma_{x'}} \tag{5}$$

$$\eta_u^s = \frac{A_{u,0}^s + \alpha}{A_{u,0} + S\alpha} \qquad \pi_u^t = \frac{A_{u,1}^t + \alpha}{A_{u,1} + T\alpha} \tag{6}$$

And a user's overall interest topics including serious topics and unserious topics is expressed by $[\lambda_u^1 \cdot \pi_u, \lambda_u^0 \cdot \eta_u]$ which is a $T + S$ dimensional multinomial distribution of topics.

To infer a new post, we use the counters of topic and indicator variable assignments in the training corpus as our prior counters and then run Gibbs sampling on the new post. After convergence, the parameters of the new post and its author can be calculated by equations 5 and 6 with the new counters.

4. Experiments and Analysis

4.1. Data Set and Experiment Setup

To evaluate our proposed model, we use forum threads from Tianya[1], a popular online forum site in China. We crawled the high-ranking threads under different boards including "Car", "IT", "Economy", "Basketball", "Travel" and so on. After tokenizing the data, all stop words and words occurring less than 10 times are removed. We also deleted users who have fewer than 3 posts. The detailed statistics of the processed forum dataset are summarized in Table 3.

Table 3: Statistics of Tianya dataset.

#Threads	#Users	#Response post
1487	5265	27950
#Words	#Response/Thread	#Response/User
37683	18.8	5.3

We compare our model with LDA [8], Author-Topic (AT) model [2], Twitter-LDA [17] and Twitter-User (TU) model [18] on this dataset in terms of topic

[1] http://bbs.tianya.cn/

quality and user interest profile. Author-Topic model represents each user by a probability distribution over interest topics and assumes a response post as a mixture of topics which are generated from its user's interest topic distribution, so it can extract user interests and forum topics simultaneously. Based on Author-Topic model, Twitter-User model further introduces a global Bernoulli distribution to decide whether a response post is related to its user's interests. Notice that in [18] the occurrences of *retweet*, *reply*, *link* and *tag* are also leveraged as features to decide the relevance of a tweet. But in online forums, these information is not available, so we use the TU model in [18] without any extra features. As for LDA and Twitter-LDA, they can not derive user interest directly, so we first train it on all response posts and then aggregate those response posts authored by the same user into a interest topic distribution. The Author-Topic model can be viewed as aggregating response posts for a user before topic modeling, and LDA and Twitter-LDA are to aggregate response posts for a user after topic modeling.

For all methods, we fix the Dirichlet priors α, β to be 0.1 and 0.01, respectively, since LDA with weak priors performs better in short texts. We also set Beta priors $\gamma_0 = 1$ and $\gamma_1 = 5$ for Forum-LDA, Twitter-LDA and Twitter-User after several trials. In all the methods, Gibbs sampling is run for 1,000 iterations to guarantee convergence. The number of serious interest topics T and unserious interest topics S are empirically set to be 20 and 1, respectively.

4.2. Topic Evaluation

We first evaluate the quality of discovered topics on Tianya dataset using Forum-LDA. Table 4 shows four discovered serious interest topics (*family*, *economy*, *basketball* and *football*), as well as unserious interest topic, with each topic represented by the 10 most probable words. The original examples are in Chinese, so we translated the Chinese to English here. As we can notice, the probable words within each topic are semantic consistent with each other. Moreover, the unserious interest topic captures common meaningless words in online forums, including joking (e.g. hahaha) and simple praise (e.g. good,

Table 4: Examples of topics discovered from Tianya dataset by Forum-LDA.

Topic	Top words
Family	children, husband, mother-in-law, parent, son, wife, marriage, man, mother, house
Economy	China, economy, nation, real estate, house price, development, government, market, city, America
Basketball	Kobe, Warriors, James, Jordan, player, Cavaliers, Curry, history, defense, data
Football	player, Evergrande, match, football, team, China team, coach, fans, level, national team
Unserious interest	not bad, up, support, follow, hahaha, thank, study, good post, post, mark

study and thank).

In order for more comprehensive analysis, we evaluate the topics generated by each method with the metric of *topic coherence* [25], which has been shown to effectively correlate with human judgement [26, 27]. For this, we use the Point-wise Mutual Information (PMI) score to calculate topic coherence. Given a topic t represented by its top N probable words $w_1, w_2, ..., w_N$, the PMI score of t is:

$$PMI(t) = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (7)$$

where $p(w_i)$ is the probability that word w_i appears in a document, and $p(w_i, w_j)$ is the probability that words w_i and w_j appear in the same document. These probabilities are estimated from a reference corpus of 300k Wikipedia articles. The overall topic coherence for each method is the averaged PMI score over all learned topics. A higher topic coherence indicates the better learned topic. Table 5 reports the topic coherence of all methods, where the number of top words N ranges from 5 to 20. As we can see, Forum-LDA achieves the highest coher-

Table 5: Topic coherence on the top N words.

N	5	10	20
LDA	2.437	2.236	2.219
AT	2.335	2.187	2.301
Twitter-LDA	2.450	2.285	2.348
TU	2.522	2.376	2.237
Forum-LDA	**2.701**	**2.471**	**2.411**

ence score across all settings. Twitter-LDA outperforms LDA and Author-Topic model slightly because its constraint on topic number improves the compactness of the derived topics. Twitter-User model also improves the topic coherence by filtering out the interest-unrelated response posts. The superior performance of Forum-LDA as compared to Twitter-LDA and Twitter-User model is in accordance with our understanding that the separation of serious interest topics and unserious interest topics, the constraint on topic number and the relationships between root posts and response posts can help enhance the quality of topics.

To further examine how Forum-LDA improves topic coherence, we also compare the topics learned by different methods semantically. Table 6 shows an example of similar topics learned by different methods. Obviously, this is a topic about *Car*s. The red tags are the words considered to be less correlated to this topic. We can see that the top words learned by Forum-LDA model is more relevant to the *Cars* topic than those learned by other methods. Because there is no explicit distinction between relevant and irrelevant response posts, the topics learned by LDA and Author-Topic model are mixed with some unrelated words such as "mark" and "sure". On the contrary, by distinguishing serious interest topics and unserious interest topics, Forum-LDA is able to identify irrelevant response posts and learn serious topics and unserious topics separately. These results further demonstrate that Forum-LDA can learn more semantic and coherent topics from online forums.

Table 6: An example of topics discovered from different models.

Model	Top words
Forum -LDA	Volkswagen, Japan, car, engine, drive, driver, domestic, technique, high rate, quality, power, Manual Transmission, Toyota, BMW, Mazda
TU	technique, Volkswagen, Manual Transmission, battery, driver, BYD, Mazda, automation, engine, feel, clutch, not bad, car, drive, kilometer
Twitter -LDA	Volkswagen, Manual Transmission, engine, drive, technique, power, car, Toyota, domestic, Mazda, USA, driver, BMW, clutch, lane change
AT	drive, Manual Transmission, driver, buy, space, Volkswagen, power, high rate, kilometer, acceleration, BYD, not bad, sure, only, clutch,
LDA	mark, drive, driver, Volkswagen, car owner, really, lane change, thank, condition, brake, high rate, speed, overtake, clutch, accident

4.3. Post Identification

Forum-LDA can be used to identify serious and unserious response posts according to the switching variable x. We treat this as a classification problem and evaluate by accuracy. To classify a response post, we treat the assignment of switching variable x after Gibbs sampling as its predicted label, that is $x = 1$ for serious post and $x = 0$ for unserious post. For comparison, we use Twitter-User model as baseline since it also has a switching variable indicating whether a response post is relevant to its user's interest. We manually annotate 400 serious posts and 400 unserious posts for evaluation. The classification results are summarized in Table 7. As we can see, Forum-LDA achieves significant improvements over Twitter-User model in both serious and unserious post classifications. The superiority demonstrates that by learning the relevance be-

Table 7: Accuracy of serious and unserious post classifications.

Method	Serious	Unserious
TU	0.738	0.685
Forum-LDA	**0.805**	**0.933**

tween root posts and response posts, the Forum-LDA model are more effective to separate serious and unserious response posts.

4.4. User Identification

Since Forum-LDA explicitly distinguish serious interest topics and unserious interest topics, it can also be used to identify serious and unserious users. This identification task is very useful as many applications such as recommendation and opinion mining are more interested in serious users rather than unserious users. In our Forum-LDA, we can learn the parameter λ_u for each user u, which represents the probability that user u publish serious and relevant response posts. Then users with high λ_u values can be considered as serious, and users with minor λ_u values can be identified as unserious. In order to evaluate the performance of this identification task, we first examine the statistics behavior of these two user sets identified by Forum-LDA. The mean value for post number and average post length for these two classes of users are summarized in Table 8. As we know, unserious user tend to publish more but shorter posts than serious users. Because while serious user usually spend more time in thinking and editing their response posts to clearly express their opinions, unserious users may just write several words to mark their appearance or agreement so they can publish more post in less time. As can be seen in Table 8, the discovered unserious users edit much more posts but use much less words in each post as we expected.

We further examine by treating user identification as a retrieval problem. We rank users according to their λ_u and use the first and last 100 users for evaluation of serious and unserious user identification, respectively. We ask

Table 8: Mean Value of average post length and number of post for the discovered serious and unserious user sets.

	#Word/Post	#Post
Serious User	33.0	5.8
Unserious User	7.2	21.3

Table 9: Precision for serious and unserious user identification.

	Method	P@10	P@25	P@50	P@100
Serious	TU	0.90	0.96	0.94	0.96
	Forum-LDA	**1.00**	**1.00**	**1.00**	**0.98**
Unserious	TU	1.00	0.92	0.78	0.62
	Forum-LDA	1.00	**1.00**	**0.95**	**0.92**

two graduate students to label the users after seeing user's response posts. We also choose Twitter-User model as baseline in which users are ranked according to the ratio of identified irrelevant posts. The evaluation results are reported in Table 9. In serious user identification, Forum-LDA obtains almost perfect performance and Twitter-User model is competitive. But in unserious user identification, Forum-LDA achieves significant improvements against Twitter-User model. These results indicate that the user-specific parameter λ_u learned by Forum-LDA is more effective than the global one learned by Twitter-User model and that the relationships between root posts and response posts is helpful to identify unserious or serious users correctly.

We demonstrate some typical posts of a serious user and an unserious user identified by our model in Table 10. For privacy, users' names are kept anonymous. User's Bernoulli distribution parameter λ_u is provided in parentheses, and the representative words of interest topic are marked in blue. From Table 10, we can find that the discovered serious user with high λ_u value is interested in topic *housing* and published many relevant response posts, while the unserious user with minor λ_u value just publishes some meaningless words or simple

Table 10: Typical posts of an unserious user and a serious user identified by Forum-LDA.

Unserious user ($\lambda_u = 0.240$)
Read.
Give you a like.
You are so great, I am going to follow you.
Thanks for your post!
Serious user ($\lambda_u = 0.915$)
House price still depends on supply and demand.
At that time, affordable housing in Shanghai was cheaper than normal housing by just million yuan. This actually disrupted the market.
In such a stable and prosperous era, the house price will only keep rising.

praises. These results further demonstrate the effectiveness of Forum-LDA in identifying unserious or serious users.

4.5. User Interest Discovery

To examine user interest discovery performance of our model and the others, we represent user interest as a multinomial distribution over interest words instead of interest topics. The probability of word w related to user u is calculated as:

$$prob(w) = \frac{\sum_t N_t^w \pi_{u,t} \phi_{t,w} + \beta}{\sum_{w'} \sum_t N_t^{w'} \pi_{t,z} \phi_{t,w} + V\beta} \quad (8)$$

where each occurrence of w with serious topic assignment t is weighted by the product of user's topic preference in t and topic t's preference in word w. The top 10 interest words of two representative users are shown in Table 11, where the red tags are the words considered to be less correlated to the interest. Based on Table 11, it is clear that User 1 is interested in *cooking,* while User 2 focuses on the topic of *stock*. In addition to serious posts, users may also publish some

unserious response posts, which is common in online forums. From Table 11 we can see that because of the frequent occurrence of irrelevant response posts, the top interest words discovered by LDA, Author-Topic model and Twitter-LDA are less correlated and mixed with some meaningless words. However, Forum-LDA can discover representative words of users' interest with its ability to explicitly identify irrelevant response post. Twitter-User model can not get rid of the frequent meaningless words as well as Forum-LDA. We consider this is because the lack of extra features which is available in tweets in [18] makes it less effective to identify irrelevant response post through a global Bernoulli distribution. But in Forum-LDA, the relevance between root posts and response posts can be utilized and a user-specific Bernoulli distribution is more flexible than a global one. These results further show the advantage of our model in user interest profiling.

5. Conclusions

In this paper, we propose a novel topic model, called Forum-LDA to model the generative process of root post, relevant and irrelevant response posts in the same thread jointly. In Forum-LDA, a user's interests are divided into serious interest topics and unserious interest topics, and the topic of relevant response post is jointly determined by its author's serious interest topics and the root post's topics, while the topic of irrelevant response post is only determined by its author's unserious interest topics. By incorporating root posts information and their relationships with response posts, Forum-LDA can not only identify irrelevant response posts, learn more semantic and coherent forum topics and user interests, but also identify serious and unserious users. The empirical experiment results on real forum dataset demonstrate the advantage of Forum-LDA in tasks such as forum topics learning, unserious post/user identification and user interest discovery.

While our results are encouraging, there is still room for improvement. One limitation of Forum-LDA is the constraint on topic number. Some longer re-

Table 11: Top interest words of two representative users.

User 1	LDA	children, tasty, hahaha, post, thing, Japan, mark, surface, go home, want
	AT	advise, yeast, tasty, last year, dough, follow, relation, machine, gluten, second floor
	Twitter -LDA	children, egg, post, go home, like, machine, dough, cake, want, sauce
	TU	time, cook, things, children, egg, no way, expensive, tasty, cake, want
	Forum -LDA	tasty, egg, children, time, gastronome, cook, no way, dough, sauce, cake
User 2	LDA	up, stock, share, follow, post, object, recommendation, lever, quotation, cost,
	AT	up, communication, share, stock, feel, object, follow, post, lever, cost
	Twitter -LDA	stock, communication, up, follow, object, share, recommendation, cost, quotation, lever
	TU	communication, up, share, recommendation, stock, many years, follow, quotation, thinking, object
	Forum -LDA	stock, quotation, communication, operation, end, cost, object, follow, Shipping space, position

sponse posts may have multiple topics, not limited to only one topic. In the future work, we tend to relax this constraint and determine the number of topics through a separate learning process. We also plan to incorporate the relationships between response posts, such as the context of response posts and the reply relationships, to accurately capture forum topics and user interests.

Acknowledgment

This work was supported by the Natural Science Foundation of Guangdong Province [grant number 2015A030313125]

References

[1] Y. Ding, J. Jiang, Q. Diao, A Unified Topic-Style Model for Online Discussions, in: Joint Workshop on Social Dynamics and Personal Attributes in Social Media, 33–41, 2014.

[2] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers, P. Smyth, The Author-Topic Model for Authors and Documents, in: UAI, 487–494, 2004.

[3] V. Cheng, C. Li, Linked Topic and Interest Model for Web Forums, in: WIIAT, 279–284, 2008.

[4] C. Li, W. K. Cheung, Y. Ye, X. Zhang, D. Chu, X. Li, The Author-Topic-Community model for author interest profiling and community discovery, Knowledge and Information System 44 (2) (2015) 359–383.

[5] H. Wu, J. Bu, C. Chen, C. Wang, G. Qiu, L. Zhang, J. Shen, Modeling Dynamic Multi-Topic Discussions in Online Forums, in: AAAI, 1455–1460, 2010.

[6] Z. Qu, Y. Liu, User Participation Prediction in Online Forums, in: EACL, 367–376, 2012.

[7] T. Hofmann, Probabilistic Latent Semantic Indexing, in: SIGIR, 50–57, 1999.

[8] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993–1022.

[9] T. L. Griffiths, M. Steyvers, Finding scientific topics, Proceedings of the National academy of Sciences 101 (suppl 1) (2004) 5228–5235.

[10] Y. Jo, A. H. Oh, Aspect and sentiment unification model for online review analysis, in: WSDM, 815–824, 2011.

[11] H. Xu, F. Zhang, W. Wang, Implicit feature identification in Chinese reviews using explicit topic mining model, Knowl.-Based Syst. 76 (2015) 166–175, doi:10.1016/j.knosys.2014.12.012, URL http://dx.doi.org/10.1016/j.knosys.2014.12.012.

[12] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, X. Lian, TIARA: Interactive, Topic-Based Visual Text Summarization and Analysis, ACM TIST 3 (2) (2012) 25.

[13] Y. Wang, E. Agichtein, M. Benzi, TM-LDA: efficient online modeling of latent topic transitions in social media, in: SIGKDD, 123–131, 2012.

[14] J. Li, M. Liao, W. Gao, Y. He, K. Wong, Topic Extraction from Microblog Posts Using Conversation Structures, in: ACL, 2114–2123, 2016.

[15] P. Zhang, H. Gu, M. Gartrell, T. Lu, D. Yang, X. Ding, N. Gu, Group-based Latent Dirichlet Allocation (Group-LDA): Effective audience detection for books in online social media, Knowl.-Based Syst. 105 (2016) 134–146, doi:10.1016/j.knosys.2016.05.006, URL http://dx.doi.org/10.1016/j.knosys.2016.05.006.

[16] D. Ramage, S. T. Dumais, D. J. Liebling, Characterizing Microblogs with Topic Models, in: ICWSM, 2010.

[17] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, X. Li, Comparing Twitter and Traditional Media Using Topic Models, in: ECIR, 338–349, 2011.

[18] Z. Xu, R. Lu, L. Xiang, Q. Yang, Discovering User Interest on Twitter with a Modified Author-Topic Model, in: WIIAT, 422–429, 2011.

[19] Z. Ren, J. Ma, S. Wang, Y. Liu, Summarizing web forum threads based on a latent topic propagation process, in: CIKM, 879–884, 2011.

[20] A. Ramesh, D. Goldwasser, B. Huang, H. Daume III, L. Getoor, Understanding MOOC Discussion Forums using Seeded LDA, in: 9th ACL Workshop on Innovative Use of NLP for Building Educational Applications, 2014.

[21] I. Hsiao, P. Awasthi, Topic facet modeling: semantic visual analytics for online discussion forums, in: LAK, 231–235, 2015.

[22] K. W. Lim, C. Chen, W. L. Buntine, Twitter-Network Topic Model: A Full Bayesian Treatment for Social Network and Text Modeling, in: NIPS Workshop on Topics Model: Computation, Application, and Evaluation, 2013.

[23] X. Pu, M. A. Chatti, H. Thüs, U. Schroeder, Wiki-LDA: A Mixed-Method Approach for Effective Interest Mining on Twitter Data, in: CSEDU, 426–433, 2016.

[24] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts, in: WWW, 1445–1456, 2013.

[25] D. Newman, J. H. Lau, K. Grieser, T. Baldwin, Automatic Evaluation of Topic Coherence, in: HLT-NAACL, 100–108, 2010.

[26] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing Semantic Coherence in Topic Models, in: EMNLP, 262–272, 2011.

[27] J. H. Lau, D. Newman, T. Baldwin, Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality, in: EACL, 530–539, 2014.