GLDA-FP: Gaussian LDA Model for Forward Prediction

Yunpeng Xiao[1](✉), Liangyun Liu[1], Ming Xu[2], Haohan Wang[3], and Yanbing Liu[1]

[1] Chongqing Engineering Laboratory of Internet and Information Security,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, China
xiaoyp@cqupt.edu.cn

[2] Research Institute of Information Technology, Tsinghua University,
Beijing 100084, China

[3] School of Computer Science, Language Technologies Institute,
Carnegie Mellon University, Pittsburgh, PA, USA

Abstract. In social networks, information propagation is affected by diversity factors. In this work, we study the formation of forward behavior, map into multidimensional driving mechanisms and apply the behavioral and structural features to forward prediction. Firstly, by considering the effect of behavioral interest, user activity and network influence, we propose three driving mechanisms: interest-driven, habit-driven and structure-driven. Secondly, by taking advantage of the Latent Dirichlet allocation (LDA) model in dealing with problems of polysemy and synonymy, the traditional text modeling method is improved by Gaussian distribution and applied to user interest, activity and influence modeling. In this way, the user topic distribution for each dimension can be obtained regardless of whether the word is discrete or continuous. Moreover, the model can be extended using the pre-discretizing method which can help LDA detect the topic evolution automatically. By introducing time information, we can dynamically monitor user activity and mine the hidden behavioral habit. Finally, a novel model, Gaussian LDA, for forward prediction is proposed. The experimental results indicate that the model not only mine user latent interest, but also improve forward prediction performance effectively.

Keywords: Multidimensional driving mechanisms · Forward prediction LDA

1 Introduction

From the perspective of information propagation, forwarding is viewed as an atomic behavior. Published messages are visible to the followers, and a follower can quickly share information that he or she is interested in. Considering the direct driving force on information diffusion, forward prediction has been applied in many fields [1, 2]. The relevant research can help us explore the direction of information dissemination [3], and has a positive significance to public opinion control [4]. Although improved

F. Y. L. Chin et al. (Eds.): BIGDATA 2018, LNCS 10968, pp. 124–139, 2018.
https://doi.org/10.1007/978-3-319-94301-5_10

methods have achieved positive results in current research on forward prediction, some challenges still remain.

On the one hand, user behavior in social networks is caused by complex factors [5]. Users are the core media of online social network information dissemination and directly affect the breadth and depth of information dissemination. Although the current research concerns the impact of user attributes on information forwarding, it only considers the basic attributes, such as the number of fans or friends [6], and ignores the intrinsic mechanisms [7] affecting forward behavior such as interests and habits [8]. Moreover, user interests tend to be multidimensional. This may make users participate in different social actions and be influenced by different users. However, there are few studies that consider the multidimensional interests of users. We need to explore more in the future research.

On the other hand, a lot of methods for forward prediction in social networks consider only static features and attributes [9, 10], and few works take time factor into consideration [11]. Not only nodes and edges are changing with time, the information forwarding behavior also change with time. Incorporating time factor into forward prediction methods would be promising.

In order to analyze the complexity of user forward behavior, it is mapped into multiple mechanisms. Both internal driving mechanisms such as interests, habits and external driving mechanism of network structure are considered. By introducing time information, we propose a model to dynamically monitor user forward behavior. To verify our proposed model, we choose real data of the Sina Weibo for evaluation. Experimental results indicate that the model not only mine user latent topic in multiple dimensions, but also improve forward prediction performance.

Our contribution can be summarized as follows:

- In order to analyze the complexity of user forward behavior, it is mapped into multiple mechanisms: interest-driven, habit-driven and structure-driven. By analyzing and quantifying these driving mechanisms, we can effectively predict user forward behavior.
- Owing to the continuity of some user attributes, the traditional LDA text modeling method is improved by Gaussian distribution and applied to user interest, activity and influence modeling. In this way, the user topic distribution for each dimension can be obtained regardless of whether the word is discrete or continuous.
- Given the insufficiency of current model consider only static features and the advantage of pre-discretizing method on helping LDA detect the topic evolution automatically. Our model can be extended using the pre-discretizing method. By introducing time information, we can dynamically monitor user activity and mine the hidden behavioral habit.

The remainder of this paper is organized as follows. Section 2 introduces related work. Section 3 formulates the problem and gives the necessary definitions. Section 4 explains the proposed model and describes the learning algorithm. Section 5 presents and analyzes the experimental results. Finally, Sect. 6 concludes the paper.

2 Related Work

In online social networks, information dissemination is mainly depended on forward behavior. Forward prediction is achieved by learning the user interests and behavior patterns. According to different assumptions, we structure the discussion of related work onto two broad previously mentioned categories: user behavior and interest modeling, dynamic modeling.

Forward Prediction with User Behavior and Interest Models. There are some prior works that focused on predicted forwarding by similar interests [12, 13]. These approaches treat forwarding as the way people interact with the messages. It is critical in understanding user behavior patterns and modeling user interest. Qiu et al. [14] proposed an LDA-based behavior-topic model which jointly models user topic interests and behavioral patterns. Bin et al. [15] proposed two novel Bayesian models which allow the prediction of future behavior and user interest in a wide range of practical applications. Comarela et al. [16] studied factors that influence a users' response, and found that the previous behavior of user, the freshness of information, the length of message could affect the users' response. However, most of the existing work is difficult to take into account the complex drivers of user behavior and neglected the intrinsic mechanisms such as habits.

Modeling and Predicting Forward Behavior Dynamically. Many studies propose dynamic modeling of user behavior. Ahmed et al. [17] proposed a time-varying model. They assumed that user actions are fully exchangeable, and that users' interest are not fixed over time. The paper divided user actions into several epochs based on the time stamp of the action and modeled user action inside each epoch using LDA. Liu et al. [18] proposed a fully dynamic topic community model to capture the time-evolving latent structures in such social streams. Moreover, some researches [19–21] assumed data in similar space are exchangeable and effectively capture the dynamics of topics in message. Zhao et al. [22] proposed a dynamic user clustering topic model. The model adaptively tracks changes of each user's time-varying topic distribution based both on the short texts the user posts during a given time period and on the previously estimated distribution. Most of the methods implement dynamic modeling by quantifying user action or interest.

This paper models user interest and behavior by using GLDA which integrate the Gaussian distribution into LDA. The internal driving mechanisms and external driving mechanisms are jointed into our model. Meanwhile, time factor is introduced by pre-discretizing method, which can help GLDA detect the topic evolution dynamically. Since then, we can obtain user interest and mine the hidden behavioral habit, as well as predict forward behavior dynamically.

3 Problem Definition

3.1 Related Definitions

We use $G = (V, E)$ to denote the structure of a social network, where V is the set of all users and E is an $N \times N$ matrix, with each element $e_{m,n} = 0$ or 1 indicating whether user v_m has a link to user v_n. The cardinality $|V| = N$ is used to denote the total number of whole network users. For predicting the forward behavior, some basic concepts and related definitions are introduced.

Definition 1. Interest following vector $\vec{e}_m^{(a)} = \left[e_{m,1}^{(a)}, e_{m,2}^{(a)}, \ldots, e_{m,N_m}^{(a)}\right]$.

We hold the view that users are more likely to follow a user they are interested in. Therefore, the user following behavior is used to define the interest following vector. $e_{m,n}^{(a)}(n = 1, 2, \ldots, N_m)$ is a user followed by user v_m, referred to here as a followed user. N_m is the number of followed users.

Definition 2. Interest interacting vector $\vec{e}_m^{(i)} = \left[e_{m,1}^{(i)}, e_{m,2}^{(i)}, \ldots, e_{m,N'_m}^{(i)}\right]$.

The following relationship only indicates the possibility of interaction between users and reflects the static user interests. By analyzing the historical interaction, we can mine active user interests. $e_{m,n}^{(i)}$ $(n = 1, 2, \ldots, N'_m)$ is a user interacted with user v_m, referred to here as an interacted user. N'_m is the number of interacted users.

Definition 3. Interest-driven vector $I(v_m) = \left[e_{m,1}^{(a)}, \ldots, e_{m,N_m}^{(a)}, e_{m,1}^{(i)}, \ldots, e_{m,N'_m}^{(i)}\right]$.

The interest-driven vector is referred to as a user interest document, which can also be expressed as the superposition of interest followed users and interacted users. Each followed user or interacted user can be referred to here as a behavioral user.

Definition 4. Habit-driven vector $A(v_m, t) = [x_{m,t,1}, x_{m,t,2}]$.

The activity is divided into post activity $x_{m,t,1}$ and forward activity $x_{m,t,2}$. Considering the characteristics of user daily routine, we divide a day into four six-hour slices and map the activity related attributes into multiple time slices, which are defined as:

$$\begin{cases} x_{m,t,1} = n_{m,t}^{pos} / n^{pos} \\ x_{m,t,2} = n_{m,t}^{ret} / n^{ret} \end{cases} \tag{1}$$

Where $n_{m,t}^{pos}$ and $n_{m,t}^{ret}$ represent the average post or forward number of user v_m at time t, n^{pos} and n^{ret} are the average post or forward number per day.

Definition 5. Structural-driven vector $S(v_m) = [d_{m,1}, d_{m,2}, d_{m,3}]$.

Network provides substrate for information propagation, thus forward behavior strongly depends on network structure. Based on the network influence related attributes, we can define structural-driven vector, where $d_{m,1}$, $d_{m,2}$, $d_{m,3}$ are the in-degree, out-degree, and node degree centrality respectively.

3.2 Problem Formulation

To formally formulate the problem of our research, let $G = (V, E)$ be the whole network, $B = \{(b, v, t) | v \in V\}$ represents the behavior information of all users. Firstly, the cause of forward behavior can be mapped into multidimensional vectors: I, A, S. If a user published a message at time slice t, then we can use our method to predict fans forward behavior Y. Specifically, the problem is formulated as follows:

$$\left.\begin{array}{c} G, B \to I, A, S \\ t \end{array}\right\} \Rightarrow f : (I, A, S, t) \to Y \tag{2}$$

4 Proposed Model

To solve the above problems, we propose a novel prediction model, GLDA, based on user behavior and relationships. The details of the model framework are introduced in three modules: driving mechanisms quantification, user interest, activity and influence modeling and forward prediction, as shown in Fig. 1. In the first module, related attributes are considered for driving mechanisms quantification, and multiple driven vectors are proposed to represent them. In the second module, the user topic distribution for each dimension can be obtained based on improved LDA. In the third module, using Gibbs sampling method to get the probability distribution of forward behavior, then the model can be proposed to do forward prediction.

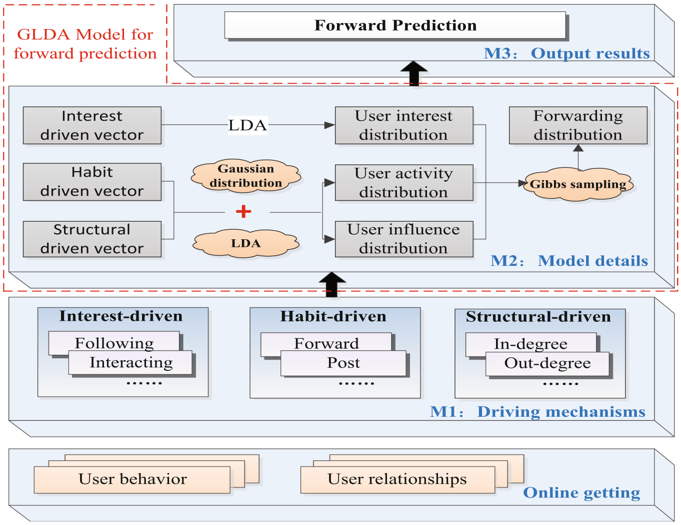


Fig. 1. Model framework.

4.1 Model Details

Given the three driven vectors defined in Sect. 3.1, the problem to be solved becomes how to incorporate those vectors into multiple prediction features and behavior modeling. This module presents the process of modeling, includes: interest-driven simulation analysis, habit-driven simulation analysis, structural-driven simulation analysis. Corresponding to different driving mechanisms, the relevant distributions of prediction features can be obtained.

Interest-Driven Simulation Analysis. User interest is reflected primarily in user behavior. We focus on the analysis of following behavior and interacting behavior. Taking advantage of the LDA topic model in dealing with polysemy and synonym problems, the traditional text modeling method is used to model user interest. Each user can be understood as the component of followed users and interacted users, which can also be expressed as its interest-driven vector. Given parameter Z as the number of interest topics, the simulated interest-driven vector generative process is:

1. For each interest topic z, draw $\vec{\xi}_z \sim Dir(\lambda)$;
2. Given the mth user v_m, in whole network G, draw $\vec{\varphi}_m \sim Dir(\alpha)$;
3. For the nth behavioral user in the mth user $e_{m,n}$:
   a. Draw an interest topic $z = z_{m,n} \sim Mult(\vec{\varphi}_m)$;
   b. Draw a behavioral user $e_{m,n} \sim Mult(\vec{\xi}_{z_{m,n}})$;
   Here, $Dir(.), Mult(.)$ denotes Dirichlet distribution and Multinomial distribution. The graphic model is shown in Fig. 2 and the symbols are described in Table 1.

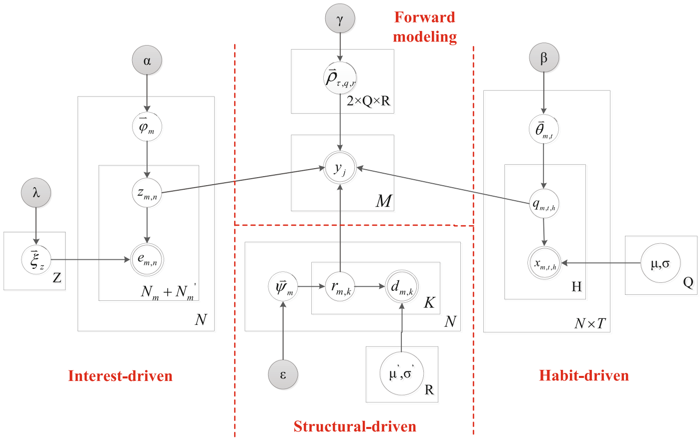


Fig. 2. Graphic model.

Actually, the aim of user interest modeling is to compute Multinomial distributions $\Phi = \left[\vec{\varphi}_1, \vec{\varphi}_2, \ldots, \vec{\varphi}_N\right]$ and $\Sigma = \left[\vec{\xi}_1, \vec{\xi}_2, \ldots, \vec{\xi}_Z\right]$. Owing to the coupling of Φ and Σ, we

Table 1. Description of symbols in graphic model.

Symbols	Descriptions	Symbols	Descriptions
$\alpha, \beta, \lambda, \gamma, \varepsilon$	Dirichlet priors	N	Number of the users in G
N_m, N'_m	Number of followed or interacted users about v_m	Z, Q, R	Number of interest or activity or influence topics
T	Number of time slices	$\vec{\varphi}_m$	Interest distribution of v_m
$\vec{\xi}_z$	Behavioral distribution of z	$e_{m,n}$	The *nth* behavioral user of v_m
$\vec{\psi}_m$	Influence distribution of v_m	$z_{m,n}$	Interest topic assigned to $e_{m,n}$
$\vec{\theta}_{m,t}$	Activity distribution of v_m at time slice t	$d_{m,k}$	The *kth* influence related attribute of v_m
$q_{m,t,h}$	Activity topic assigned to $x_{m,t,h}$	$r_{m,k}$	Influence topic assigned to $d_{m,k}$
$x_{m,t,h}$	The *hth* activity related attribute of v_m at time slice t	H, K	Number of activity or influence related attributes
$\mu, \sigma, \mu', \sigma'$	Parameters of Gaussian distributions	$\vec{\rho}_{\tau,q,r}$	Forward behavior distribution of multiple prediction features
y_j	The *jth* forward behavior to be predicted	M	Number of forward behaviors to be predicted

cannot compute them directly and Gibbs sampling [23] is applied to indirectly get Φ and Σ. The principle of Gibbs sampling in terms of extracting topic z_i of behavior user e_i is as follows:

$$\begin{aligned} p\left(z_i = z|\vec{z}_{-i}, E\right) &\propto p\left(z_i = z, e_i = e|\vec{z}_{-i}, E\right) = \widehat{\varphi}_{m,z} \times \widehat{\xi}_{z,e} \\ &= \frac{n^{(z)}_{m,-i} + \alpha}{\sum_{z=1}^{Z} n^{(z)}_{m,\neg i} + \alpha} \times \frac{n^{(e)}_{z,-i} + \beta}{\sum_{e=1}^{N} n^{(e)}_{z,\neg i} + \beta} \end{aligned} \tag{3}$$

Where $\vec{z}_{\neg i}$ represents the topic of behavioral users except for the current behavioral user; $\vec{e}_{\neg i}$ represents behavioral users except for the current behavioral user; $n^{(e)}_{z,\neg i}$ is the number of behavioral user e assigned to interest topic z except for the current behavioral user; and $n^{(z)}_{m,\neg i}$ is the number of interest topic z assigned to user v_m except for the current behavioral user. When the sampling converges, Φ and Σ can be obtained.

Habit-Driven Simulation Analysis. The user behavioral habit can be analyzed based on historical behavior. Here, we focus on post behavior and forward behavior. Considering the dynamics of user behavioral habit, the past behavioral data are pre-discretized to get habit-driven vector. In other words, a user at every time slice is regarded as a document and the activity related attributes are regarded as words. By introducing activity as topics, we can mine potential user behavioral habits.

Unlike the value of $e_{m,n}$, the activity related attributes are both continuous. Owing to the useless for dealing with continuous attributes modeling, Gaussian distribution is used to replace Multinomial distribution in standard LDA. In the improved model, the activity related attributes obey the following Gaussian distribution:

$$f(x_{m,t,h}; \mu_{q,h}, \sigma_{q,h}) = \frac{1}{\sqrt{2\pi}\sigma_{q,h}} \exp\left[-(x_{m,t,h} - \mu_{q,h})^2 / 2\sigma_{q,h}^2\right] \tag{4}$$

Where $x_{m,t,h}$ is the hth attribute of user v_m at time slice t, and $\mu_{q,h}, \sigma_{q,h}$ are parameters of Gaussian distribution that $x_{m,t,h}$ obeys. The simulated habit-driven vector generative process can be described as follows:

1. Given the mth user v_m at any time slice t, draw $\vec{\theta}_{m,t} \sim Dir(\beta)$;
2. For the hth activity related attribute of the mth user $x_{m,t,h}$:

   a. Draw an activity topic $q = q_{m,t,h} \sim Mult(\vec{\theta}_{m,t})$;
   b. Draw an attribute value $x_{m,t,h} \sim N(\mu_{q_{m,t,h},h}, \sigma_{q_{m,t,h},h})$;

Where $N(.)$ denotes Gaussian distribution. The purpose of user activity modeling is to learn the distribution set $\Pi = [(\mu_{1,h}, \sigma_{1,h}), (\mu_{2,h}, \sigma_{2,h}), \ldots, (\mu_{Q,h}, \sigma_{Q,h})]$ $(h \in [1, H])$ and $\Theta = [\vec{\theta}_{1,t}, \vec{\theta}_{2,t}, \ldots, \vec{\theta}_{N,t}]$ $(t \in T)$. Owing to the existence of hidden variables, the EM algorithm [24] is used to estimate model parameters. E step computes the responsiveness of topic to attribute according to the current model parameters:

$$\chi_{m,t,h,q} = P(q|v_m, t, x_{m,t,h}) = P(x_{m,t,h}|q) \times P(q|v_m, t) = \frac{f(x_{m,t,h}; \mu_{q,h}, \sigma_{q,h})\theta_{m,t,q}}{\sum_{q'=1}^{Q} f(x_{m,t,h}; \mu_{q',h}, \sigma_{q',h})\theta_{m,t,q'}} \tag{5}$$

M step updates the model parameters for the new round of iteration:

$$\mu_{q,h} = \frac{\sum_{m=1}^{N} \sum_{t=1}^{T} \chi_{m,t,h,q} * x_{m,t,h}}{\sum_{m=1}^{N} \sum_{t=1}^{T} \chi_{m,t,h,q}}, \quad \sigma_{q,h} = \sqrt{\frac{\sum_{m=1}^{N} \sum_{t=1}^{T} \chi_{m,t,h,q}(x_{m,t,h} - \mu_{q,h})^2}{\sum_{m=1}^{N} \sum_{t=1}^{T} \chi_{m,t,h,q}}} \tag{6}$$

$$\theta_{m,t,q} = \frac{1}{H} \sum_{h=1}^{H} \chi_{m,t,h,q} \tag{7}$$

Where $\chi_{m,t,h,q}$ is the responsiveness of activity topic q to the attribute $x_{m,t,h}$. $\theta_{m,t,q}$ denotes the probability of user v_m assigned to activity topic q at time slice t. Repeat the above two steps until convergence, Θ and Π can be obtained.

Structural-Driven Simulation Analysis. The network structure contains many user attributes, such as in-degree, out-degree, and other attributes, which can be expressed as structural-driven vector. Based on it, we can classify users into clusters. Each cluster can be regarded as an influence role that users play. Each influence role has a set of parameters of distribution that the influence related attributes conform to. Similar with the previous Section, we also use Gaussian distribution. If user v_m play influence role r, its kth attribute $d_{m,k}$ conforms to:

$$f\left(d_{m,k}; \mu'_{r,k}, \sigma'_{r,k}\right) = \frac{1}{\sqrt{2\pi}\sigma'_{r,k}} \exp\left[-\left(d_{m,k} - \mu'_{r,k}\right)^2 / 2\sigma'^2_{r,k}\right] \tag{8}$$

Where $\mu'_{r,k}$ and $\sigma'_{r,k}$ are parameters of Gaussian distribution that $d_{m,k}$ obeys. The simulated structural-driven vector generative process is as follows:

1. Given the *mth* user v_m, in the whole network, draw $\vec{\psi}_m \sim Dir(\varepsilon)$;
2. For the *kth* influence related attribute of the *mth* user $d_{m,k}$:
   a. Draw an influence topic $r = r_{m,k} \sim Mult(\vec{\psi}_m)$;
   b. Draw an attribute value $d_{m,k} \sim N(\mu'_{r_{m,k},k}, \sigma'_{r_{m,k},k})$;

Our goal is to learn distributions $\Pi' = [(\mu'_{1,k}, \sigma'_{1,k}), (\mu'_{2,k}, \sigma'_{2,k}), \ldots, (\mu'_{R,k}, \sigma'_{R,k})]$ $(k \in [1, K])$ and $\Psi = [\vec{\psi}_1, \vec{\psi}_2, \ldots, \vec{\psi}_N]$. And EM algorithm is also used to estimate model parameters. E step computes the responsiveness as:

$$\chi'_{m,k,r} = P\left(r|v_m, d_{m,k}\right) = P\left(d_{m,k}|r\right) \times P(r|v_m) = \frac{f(d_{m,k}; \mu'_{r,k}, \sigma'_{r,k})\psi_{m,r}}{\sum_{r'=1}^{R} f(d_{m,k}; \mu'_{r',k}, \sigma'_{r',k})\psi_{m,r'}} \tag{9}$$

M step updates the model parameters for the new round of iteration:

$$\mu'_{r,k} = \frac{\sum_{m=1}^{N} \chi'_{m,k,r} d_{m,k}}{\sum_{m=1}^{N} \chi'_{m,k,r}}, \quad \sigma'_{r,k} = \sqrt{\frac{\sum_{m=1}^{N} \chi'_{m,k,r} (d_{m,k} - \mu'_{r,k})^2}{\sum_{m=1}^{N} \chi'_{m,k,r}}} \tag{10}$$

$$\psi_{m,r} = \frac{1}{K} \sum_{k=1}^{K} \chi'_{m,k,r} \tag{11}$$

Where $\chi'_{m,k,r}$ is the responsiveness of influence role r to the attribute $d_{m,k}$. $\psi_{m,r}$ denotes the probability of user v_m assigned to influence role r. Repeat the two steps until convergence, Ψ and Π' can be obtained.

4.2 Comprehensive Forward Behavior Modeling and Prediction

Based on the previous modeling process, the relevant probability distributions of prediction features are obtained. By combining these distributions, the forward behavior distribution can be computed to predict the forward action that a user may take. Assume that $Y = \{y_1, y_2, \ldots, y_M\}$ is the behavior set to be predicted and its generative process can be described as follows:

1. Draw $\vec{\rho} \sim Dir(\gamma)$;
2. For the *jth* behavior y_j:
   a. Draw an interest topic $z_m \sim Mult(\vec{\varphi}_m)$ for post user v_m;
   b. Draw an interest topic $z_n \sim Mult(\vec{\varphi}_n)$ for fans v_n;
   c. Draw an activity topic $q = q_{n,t} \sim Mult(\vec{\theta}_{n,t})$ for fans v_n;

d. Draw a influence role $r = r_m \sim Mult(\vec{\psi}_m)$ for post user v_m;

e. Draw the behavior $y_j \sim Mult(\vec{\rho}_{\tau,q,r})$;

Where τ is an indicator function of interest, if $z_m = z_n$, $\tau = 1$, otherwise $\tau = 0$. Behavior y_j only contains two cases ($y_j = 1$ indicates establish forward action, $y_j = 0$ indicates not establish forward action), so we can use a Bernoulli distribution $\vec{\rho}_{\tau,q,r}$ to represent the probability distribution of multiple features over forward actions and the parameter $\Omega = \left[\vec{\rho}_{0,q,r}, \vec{\rho}_{1,q,r}\right]$ $(q \in [1, Q], r \in [1, R])$. By using Gibbs sampling, the principle of extracting features τ, q, r of behavior y_j is as follows:

$$\begin{aligned} p\left(\tau_j = \tau, q_j = q, r_j = r | \vec{\tau}_{\neg j}, \vec{q}_{\neg j}, \vec{r}_{\neg j}, Y\right) &\propto p\left(\tau_j = \tau, q_j = q, r_j = r, y_j = y | \vec{\tau}_{\neg j}, \vec{q}_{\neg j}, \vec{r}_{\neg j}, Y_{\neg j}\right) \\ &= p(\tau|\Phi)p(q|\Theta)p(r|\Psi)\hat{\rho}_{\tau,q,r,y} \\ &= \left(\vec{\varphi}_m \vec{\varphi}_n^T\right)\theta_{n,t}^{(q)}\psi_m^{(r)} \times \frac{n_{\tau,q,r,\neg j}^{(y)} + \gamma}{\sum_{y=0}^{1} n_{\tau,q,r,\neg j}^{(y)} + \gamma} \end{aligned} \tag{12}$$

Where $\vec{\tau}_{\neg j}, \vec{q}_{\neg j}, \vec{r}_{\neg j}$ represents the prediction features of behavioral except for the current behavior; $Y_{\neg j}$ represents behavior to be predicted except for the current behavior; $n_{\tau,q,r,\neg j}^{(y)}$ is the number of behavior y assigned to prediction features τ, q, r except for the current behavior; When the sampling converges, Ω can be obtained.

Given a message, we can predict the forward behavior based on the trained model. Firstly, get the features τ, q, r for model input by probability sampling method, and then the user's forward probability $\rho_{\tau,q,r,1}$ and non-forward probability $\rho_{\tau,q,r,0}$ are calculated according to the parameter Ω. If $\rho_{\tau,q,r,1} > \rho_{\tau,q,r,0}$, we predict the fans will forward the message, $y = 1$; Otherwise it will not, $y = 0$, formally expressed as:

$$y = \mathrm{argmax}_{\Omega}\, p(|y|\tau, q, r) \tag{13}$$

5 Experiments and Analysis

5.1 Experimental Data and Evaluation Metrics

The experimental data used in this paper is collected from Sina micro-blog, a popular social networking platforms in China. In the process of data collection, we randomly selected a user (user ID: 2312704093) as the starting point. Some users and their micro-blogs are captured based on breadth-first-search, forming a sub-network containing 49,556 users and 61,880 user relationships for the 2011/08/21-2012/02/22. The statistics of the dataset is shown in Table 2.

In this paper, Accuracy, Precision, Recall, F1-Measure, and ROC curve were used to verify the prediction results. We assumed that the forward behavior of fans is a positive example "1", and non-forward behavior is a negative example "0". Meanwhile, the dataset

Table 2. Statistics of the dataset

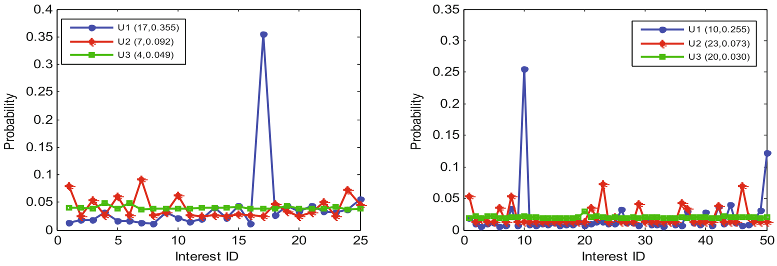
Item	Users	Relationships	Post	Forward	Review
Count	49,556	61,880	3,057,635	506,765,237	185,079,821

needs to be partitioned into training set and test set. Here, we set the proportion of training set and test set to be 8:2. The better prediction results have greater Accuracy, Precision, Recall, F1-Measure, and their ROC curves are close to the upper left corner.

5.2 Prediction Performance Analysis

In this section, the performances of our model are evaluated from three viewpoints. Firstly, we show the results of user latent interest distribution and analyze the overall interest distribution of network. Then, the impact of interest number and the proportion of training set on forward prediction can be verified. Finally, we evaluate the performance by comparing our model with other baseline methods. According to the above three viewpoints, the superiority of our model can be verified.

Firstly, the result of user latent interest distribution is analyzed. We select several representative users to show their latent interests in Fig. 3. Meanwhile, latent interest distributions in network can be shown in Fig. 4. Both in the figures, the x-axis represents the interest ID and the y-axis represents the probability value. The highest interest focus values are provided in parentheses in the legend.

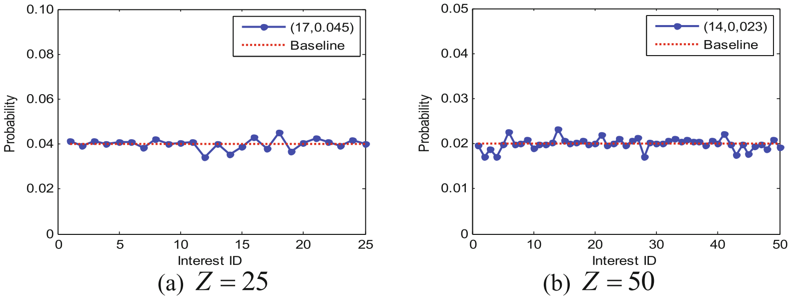


(a) $Z = 25$ (b) $Z = 50$

Fig. 3. User latent interest distributions.

As shown in Fig. 3, the distribution of each user interest is different. When latent interest number $Z = 25$, user $U1$ interest is obvious and prefer Interest $ID = 17$. The range of user $U2$ interest is relatively wide and the proportion interest of user $U3$ is average. And from Fig. 4, we can observe that the network interest distribution is uniform, although user preferences in the entire network have some differences. Next we will verify the latent interest has a driving effect on forward prediction.

Secondly, considering the excellent classification effect of LR and SVM, they are applied to the forward prediction problem and compared with our model. And the effects of the proportion of training set on forward prediction are shown in Fig. 5.

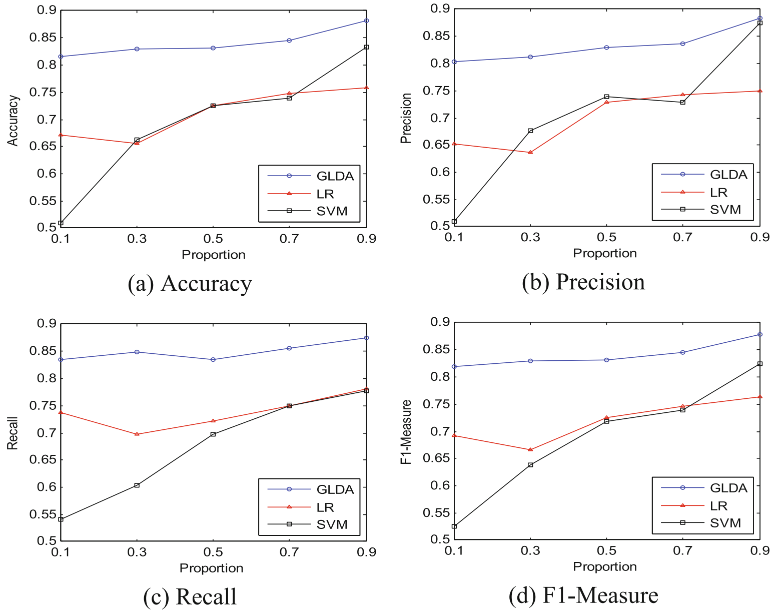


(a) $Z = 25$ (b) $Z = 50$

Fig. 4. Latent interest distributions in network.

In addition, by reducing the dimension of driving mechanism in our model, three sub-models are obtained: Sub-IA, Sub-IS, and Sub-I. Comparing our model with the sub-models, the effects of the interest topic number on each model can be shown as Fig. 6.

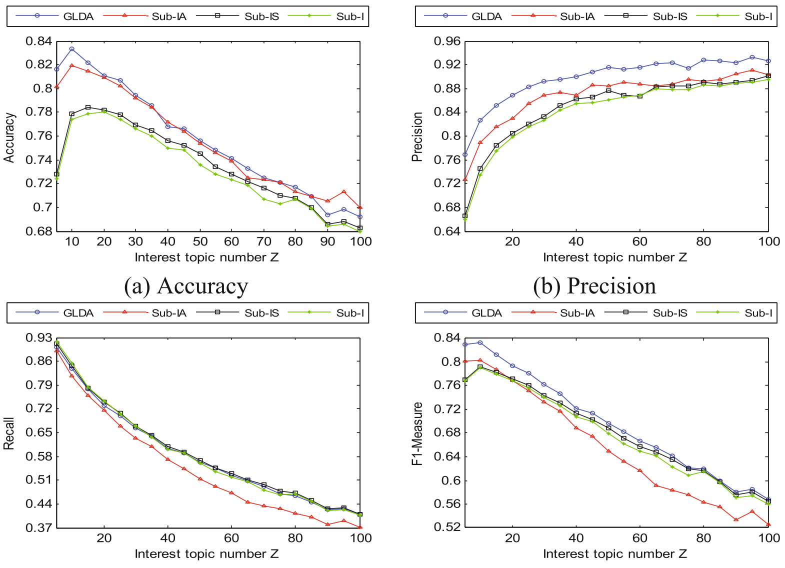
As shown in Fig. 5, the performance of GLDA is better than LR and SVM. And the performance of our model is least affected by the training set proportion. Overall, as the proportion of training set increases, the prediction effect of each method improves. From Fig. 6, we also can see our model has better prediction performance than its sub-models. It indicates that the extraction of multidimensional driven vector can improve the effect of forward prediction. With the increase of interest topic number, the Precision increases gradually, while the Recall decreases rapidly. From the change of F1-Measure, we can see that when Z is 10–20, the model performs well. In addition, the



(a) Accuracy (b) Precision

(c) Recall (d) F1-Measure

Fig. 5. Comparison of prediction effects between proposed model and classifiers.



(a) Accuracy

(b) Precision

(c) Recall

(d) F1-Measure

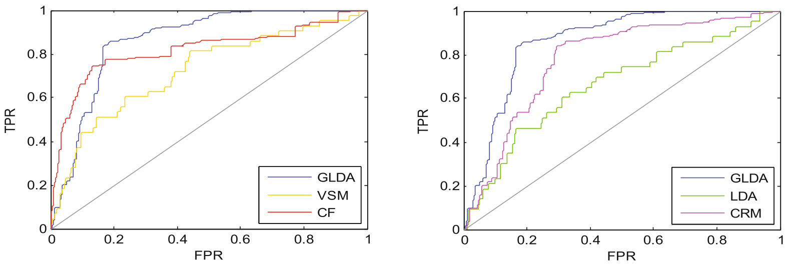
Fig. 6. Comparison of prediction effects between proposed model and sub-models.

number of activity topics and influence topics are proposed in the literature [25, 26]. It proposes to classify the activity level as inactive, generally active and very active and points out that users can be divided into three types: ordinary user, opinion leader and structural hole spanner.

Finally, the performance of our model is evaluated by comparing with some baseline methods, such as probabilistic graph model LDA [27] and CRM [6], classical forward prediction methods CF [28] and VSM [29]. The performances of them are shown in Table 3. And ROC curves comparison are shown in Fig. 7. The results show that our model plays optimal performance in Accuracy, Recall and F1-Measure compared with baseline methods. And it can be seen that the ROC curve of our proposed model is closest to the upper left corner, and the overall performance is best. Therefore, our model can improve the forward prediction performance effectively.

Table 3. Comparison between our model and baseline methods.

Methods	Accuracy	Precision	Recall	F1-Measure
LDA	0.774	0.734	0.805	0.768
CRM	0.784	0.785	0.822	0.803
CF	0.805	**0.848**	0.746	0.794
VSM	0.665	0.703	0.742	0.722
GLDA	**0.833**	0.827	**0.838**	**0.833**



(a) ROC1 comparison (b) ROC2 comparison

Fig. 7. Comparison of different methods in ROC.

6 Conclusion

In this study, a novel forward prediction model GLDA is proposed, and it can effectively predict forward behavior by analyzing user behavior and relationships. Firstly, we mapped the cause of forward behavior into three driving mechanisms: interest-driven, activity-driven and structure-driven. Secondly, the traditional LDA was improved by Gaussian distribution and applied to user interest, activity and influence modeling. Finally, the model was extended with the pre-discretizing method and we can dynamically monitor user activity and mine the hidden behavioral habit.

The experimental results showed that our model can improve forward prediction performance in comparison to other baseline methods. By studying forward prediction in social networks, we can acquire a better understanding of information propagation mechanism. And the model can provide support for public opinion management and control. In the future work, it would be intriguing to integrate nonparametric methods into our model to base parameter value choices on the data itself. It is also interesting to explore a method to alleviate the time complexity of training algorithms.

Acknowledgements. This paper is partially supported by the National 973 Key Basic Research Program of China (Grant No.2013CB329606), the National Natural Science Foundation of China (Grant No.61772098), Chongqing Science and Technology Commission Project (Grant No.cstc 2017jcyjAX0099) and Chongqing key research and development project (Grant No.cstc2017 zdcy-zdyf0299, cstc2017zdcy-zdyf0436).

References

1. Yu, H., Bai, X.F., Huang, C.Z., et al.: Prediction of users retweet times in social network. Int. J. Multimed. Ubiquit. Eng. **10**(5), 315–322 (2015)
2. Huang, D., Zhou, J., Mu, D., et al.: Retweet behavior prediction in Twitter. In: 7th International Symposium on Computational Intelligence and Design, pp. 30–33. IEEE, Hangzhou (2015)
3. Xu, Q., Su, Z., Zhang, K., et al.: Epidemic information dissemination in mobile social networks with opportunistic links. IEEE Trans. Emerg. Top. Comput. **3**(3), 399–409 (2017)

4. Yoo, E., Rand, W., Eftekhar, M., et al.: Evaluating information diffusion speed and its determinants in social media networks during humanitarian crises. J. Oper. Manag. **45**, 123–133 (2016)
5. Xiao, Y., Li, N., Xu, M., et al.: A user behavior influence model of social hotspot under implicit link. Inf. Sci. **396**, 114–126 (2017)
6. Han, Y., Tang, J.: Probabilistic community and role model for social networks. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, pp. 407–416. ACM (2015)
7. Lee, S.: What makes twitterers retweet on Twitter? Exploring the roles of intrinsic/extrinsic motivation and social capital. J. Korea Acad.-Ind. **15**(6), 3499–3511 (2014)
8. Jin, Y., Zhai, L.H.: An Investigation and Analysis of the Impact of User's Forwarding Behavior on the Quality of Information in Social Media Environment. Libr. Theor. Pract. September 2016
9. Wang, C., Li, Q., Wang, L., et al.: Incorporating message embedding into co-factor matrix factorization for retweeting prediction. In: International Joint Conference on Neural Networks, Anchorage, AK, USA, pp. 1265–1272. IEEE (2017)
10. Zhang, Y., Lyu, T., Zhang, Y.: Hierarchical community-level information diffusion modeling in social networks. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, pp. 753–762. ACM (2017)
11. Zarrinkalam, F., Kahani, M., Bagheri, E.: Mining user interests over active topics on social networks. Inf. Process. Manag. **54**(2), 339–357 (2018)
12. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, pp. 199–208. ACM (2009)
13. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, D.C., pp. 137–146. ACM (2003)
14. Qiu, M., Zhu, F., Jiang, J.: It is not just what we say, but how we say them: LDA-based behavior-topic model. In: Proceedings of the 2013 SIAM International Conference on Data Mining, pp. 794–802. Society for Industrial and Applied Mathematics (2013)
15. Bi, B., Cho, J.: Modeling a retweet network via an adaptive Bayesian approach. In: Proceedings of the 25th International Conference on World Wide Web, Canada, pp. 459–469. International World Wide Web Conferences Steering Committee (2016)
16. Comarela, G., Crovella, M., Almeida, V., et al.: Understanding factors that affect response rates in Twitter. In: Proceedings of the 23rd ACM Conference on Hypertext and Social Media, Milwaukee, Wisconsin, USA, pp. 123–132. ACM (2012)
17. Ahmed, A., Low, Y., Aly M, et al.: Scalable distributed inference of dynamic user interests for behavioral targeting. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, California, USA, pp. 114–122. ACM (2011)
18. Liu, Z., Zheng, Q., Wang, F., et al.: A dynamic nonparametric model for characterizing the topical communities in social streams. In: Proceedings of the 2014 SIAM International Conference on Data Mining, pp. 379–387. Society for Industrial and Applied Mathematics (2014)
19. Ahmed, A., Xing, E.P.: Timeline: a dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. arXiv preprint arXiv:1203.3463 (2012)

20. Ahmed, A., Xing, E.P.: Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In: Proceedings of the 2008 SIAM International Conference on Data Mining, pp. 219–230. Society for Industrial and Applied Mathematics (2008)
21. Blei, D.M., Frazier, P.I.: Distance Dependent Chinese Restaurant Processes. J. Mach. Learn. Res. **12**(1), 2461–2488 (2009)
22. Zhao, Y., Liang, S., Ren, Z., et al.: Explainable user clustering in short text streams. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, pp. 155–164. ACM (2016)
23. Zhao, F., Zhu, Y., Jin, H., et al.: A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. Future Gener. Comput. Syst. **65**, 196–206 (2016)
24. Zanetti, M., Bovolo, F., Bruzzone, L.: Rayleigh-rice mixture parameter estimation via EM algorithm for change detection in multispectral images. IEEE Trans. Image Process. **24**(12), 5004–5016 (2015)
25. Zhu, Y., Zhong, E., Pan, S.J., et al.: Predicting user activity level in social networks. In: Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, San Francisco, California, USA, pp. 159–168. ACM (2013)
26. Yang, Y., Tang, J., Leung, C.W., et al.: RAIN: social role-aware information diffusion. In: AAAI, pp. 367–373 (2015)
27. Li, L., He, J., Wang, M., et al.: Trust agent-based behavior induction in social networks. IEEE Intell. Syst. **31**(1), 24–30 (2016)
28. Jiang, B., Liang, J., Sha, Y., et al.: Retweeting behavior prediction based on one-class collaborative filtering in social networks. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, pp. 977–980. ACM (2016)
29. Waitelonis, J., Exeler, C., Sack, H.: Enabled generalized vector space model to improve document retrieval. In: NLP-DBPEDIA@ ISWC, pp. 33–44 (2015)