Information Sciences 423 (2018) 219–234

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Identifying impact of intrinsic factors on topic preferences in online social media: A nonparametric hierarchical Bayesian approach

Yezheng Liu[a,b], Jiajia Wang[a,d], Yuanchun Jiang[a,b,*], Jianshan Sun[a,b], Jennifer Shang[c]

[a] School of Management, Hefei University of Technology, Hefei, Anhui 230009, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei, Anhui 230009, China
[c] The Joseph M. Katz Graduate School of Business, University of Pittsburgh, Pittsburgh, PA 15260, USA
[d] BigData PDU, Huawei Software Technologies CO.LTD, Nanjing, Jiangsu 210012, China

ARTICLE INFO

Article history:
Received 20 October 2016
Revised 8 September 2017
Accepted 15 September 2017
Available online 19 September 2017

Keywords:
Social media
Topic preference
Intrinsic factor
Nonparametric hierarchical Bayesian model

ABSTRACT

Social media offers a new communication channel for users and affords an interactive opportunity between users and the firms about the products and the brands. Understanding what topics are important to users and the corresponding internal motivation is crucial for managers to successfully engage customers and promote business through social media. Assuming topic preference is the outcome of intrinsic factors such as gender, age and personality traits, this paper proposes an improved nonparametric hierarchical Bayesian topic (NHBT) model to investigate the multiple-to-multiple generative relationships from intrinsic factors to topic preferences. The proposed NHBT model employs a three-level generation framework based on Dirichlet process to study the impact of intrinsic factors on users topic preference. Our study of Facebook data shows that NHBT model is able to draw valuable latent topics (e.g. music band, chemical biology, cosplay) from the open social media environment, and reveal the internal motivation for users topic selection behaviors (e.g. users with low conscientiousness and high extraversion personality prefer topics about campus party). Our experiments also show that NHBT model can identify the intrinsic factors dominating topic preferences for individual users, and provide foundations to predict the intrinsic factors for new user generated contents.

1. Introduction

With the development of the social media and the fast growing number of users, online social media has become the best platform for users to communicate with each other. In 2014, the number of social media users has reached 1.87 billion and it is estimated to approximate 2.55 billion in 2018. Through online social media platform, users make friends, share interests and communicate their concerns about firms, products and brands. Online social media has been an important channel for firms to attract new customers, improve brand perception and grow their business.

* Corresponding author.
E-mail addresses: liuyezheng@hfut.edu.cn (Y. Liu), wjj5881005@126.com (J. Wang), ycjiang@hfut.edu.cn (Y. Jiang), sunjs9413@hfut.edu.cn (J. Sun), shang@katz.pitt.edu (J. Shang).

http://dx.doi.org/10.1016/j.ins.2017.09.041
0020-0255/

Because of the increasing popularity and the unprecedented volume of data available in online social media, what topics are of users interest has been an important concern for firms. Various generative probabilistic models such as LDA[2], sLDA [21] and L-LDA [29] have been proposed to summarize users topic preferences from contents published in online social media. While these models are able to predict what topics users prefer[12,14,28], little works have been done to study why users prefer the topics. Cao argued that it is ineffective or even impossible to deeply scrutinize users native behavior intention if we employ entirely web usage data to analyze user behavior [3]. Understanding the incentive reason of users online social behavior provides managers a better perspective to promote their business. It helps firms gain insight into the decisions of online social networking groups [38], increase recommendation diversity and accuracy [6], and enhance customers online purchase intention [45].

Intrinsic factors are the internal motivations for a persons actions and are widely used to explain users social networking behaviors, such as the online language usage[25,34], emotion expression [1,20], advertisement preference [16], and music selection [10]. The literature proves the correlation between the internal motivations and online social behaviors. However, they tend to focus on the correlation test, with little emphasize on the generative relationships from the intrinsic factors to extrinsic behaviors. The psychology literature maintains that extrinsic behaviors are caused by intrinsic factors such as gender, age and personality [13]. The correlation tested in the current literature does not reveal the internal motivation of user behaviors. When using social media data to model internal motivation, the current literature focuses on closed-vocabulary methods without emphasizing diversified topics in online social media. They often classify user generated contents into 64 categories (e.g. topics) by the Linguistic Inquiry and Word Count (LIWC) software, and then explore the correlation between the intrinsic factors and the fixed LIWC categories [22,27,34]. The topics discussed evolve over time and free-form languages are widely adopted. The closed-vocabulary embedded in the LIWC software obviously cannot capture the topic dynamics and may cause information loss.

In this research, we propose a hierarchical Bayesian topic model to investigate the impact of intrinsic factors on users topic preferences. Using the open-vocabulary rather than closed-vocabulary, the proposed model avoids the need of pre-defined topic number and types, and draws latent topics dynamically from online social media. It is widely documented that the influence of personality on users social networking behavior differs by both gender and age [7,23]. A specific dimension of personality is also reported to have influence on different social networking behaviors. For example, agreeableness and conscientious are positively related to the religion topic [34], and negatively related to negative emotion words [44]. Therefore, different from the existing models which associate user behavior with one intrinsic factor at a time, we assume a multiple-to-multiple generative relationship from intrinsic factors to topic selection behaviors. That is, a specific topic selection behavior is the outcome of multiple intrinsic factors [34,44] and an intrinsic factor may have influence on multiple topic selection behaviors. Because it is too restrictive and not feasible to fix the number of intrinsic factors generating a specific behavior [30,37], we employ a nonparametric approach to model the process of generating topics from intrinsic factors.

The proposed nonparametric hierarchical Bayesian topic (NHBT) model employs a three-level generation process to investigate the internal motivation of topic selection behaviors. The first level (global level) defines the global random measure, which is a draw from a standard Dirichlet process [37]. The second level (factor level) defines random measures for intrinsic factors, which are draws from a Dirichlet process, where the base measure is the random measure in global level. Finally, the third level (document level) defines random measures for each document. The base measure of the Dirichlet process in document level is a mixture of random measures in the factor level. The three-level process is iterated and we propose an inference approach under the framework of Gibbs sampling. The numerical studies based on data collected from myPersonality application, a third-party application on Facebook, show that the proposed model can find diverse topics and reveal the internal motivation regulating users topic selection behavior. The contributions of the proposed model are threefold.

(1) To the best of our knowledge, this is the first paper to investigate the generative relationships between intrinsic factors and user-generated contents based on nonparametric hierarchical Bayesian topic framework. The proposed NHBT model is consistent with the action logic of behavior and internal drives, and can capture the dynamic character of online social media.
(2) Conventionally, the relationship between intrinsic driver and user behavior is regarded as a one-to-one mapping. Our model more realistically proposes a three-level derivation process which generates each topic from multiple intrinsic factors and also correlates each intrinsic factor with more than one topic. The proposed NHBT model allows statistical strength to be shared across intrinsic factors, topics and documents.
(3) Instead of applying LIWC software to obtain the pre-defined topics, the proposed NHBT model automatically extracts latent topics from online social media. Without the restriction of the closed LIWC dictionary, the proposed model is able to obtain more meaningful topics.

The remainder of the paper is organized as follows. Section 2 surveys the related work on the automatic behavior analysis of personality and topic models for labeled data. We detail the proposed NHBT model and the inference process in Section 3. Both the quantitative and qualitative evaluation results of the proposed model are presented in Section 4. Section 5 gives the conclusions and the directions of future research.

2. Related work

We analyze the internal drivers influencing users topic selection behavior. There are two streams of literature related to our research. One is behavior analysis and personality prediction based on social media data. The other is the augmented topic analysis that incorporates metadata. In this section we review the theoretical and practical models of these research streams.

2.1. Behavior analysis and personality prediction

The literature on behavior analysis and personality prediction aims to analyze the relationships between user generated content and personality traits. Researchers have examined the word usage and personality using closed-vocabulary to process the raw and noisy user generated data, and employing LIWC software to partition social media data into 64 psychological categories. For example, one can place words like Altar, church, mosque, and christ into the religion lexicon, and count how often words in this lexicon are used by a user in order to determine if he prefers the religion topic. Pennebaker and King [27] conducted a LIWC study on word usage in various domains: diaries, manuscript abstracts, and writing assignments. They found that word usage is correlated with personality traits. For example, neurotics are likely to use negative emotion words. Introverts and those with low conscientiousness are likely to use words about distinctions. By tracking the natural speech of 96 users, Mehl et al. [22] obtained similar results that neurotic and agreeable users tend to use first-person singulars, and individuals low in openness talk more topics about social process. With the growth of social network, more and more psychologists, sociologists and computer scientists are interested in the relationships between user-generated content and personality traits. By analyzing 694 bloggers over 100,000 words, Yarkoni [44] found individuals high in agreeableness employ more positive emotion words, friend words, and family words.

Because the pre-defined dictionary cannot satisfy the requirement of the open online social media environment and the free-form languages may invalidate the LIWC software, researchers have gradually shifted their attention from closed-vocabulary to open-vocabulary study. Open-vocabulary methods do not depend on pre-defined topics and can generate new topics automatically from real data. Schwartz et al. [34] adopted this approach to extract nearly 700 million single words, multiword phrases and topics, and found more linguistic cues than closed-vocabulary method. Zhang et al. proposed a tucker deep computation model which can automatically learn features from mobile multimedia [47]. They also proposed strategies to enhance the efficiency of feature learning from large-scale data [46,48]. Based on the features extracted by open-vocabulary method, Park et al. [26] built a predictive model of personality with 66,000 Facebook users. Their experiments showed that the open-vocabulary features are more compelling than close-vocabulary.

2.2. Augmented topic models

Intrinsic factors are a type of metadata, which can be incorporated in augmented topic models. The availability of metadata has motivated various topic models to jointly model both the metadata and the text content. Metadata and text content modeling can be classified into downstream modeling and upstream modeling. The downstream models refer to topic models in which the metadata is generated given latent topics. These models follow the idea that topics generate metadata. In downstream models, each topic has a distribution not only over words but also over metadata. Examples of downstream models include those that simultaneously consider text content and metadata, e.g. references [8], timestamps [41], citations [24], and age [19]. The supervised latent Dirichlet allocation (sLDA) model is a typical downstream model in which the metadata is generated from empirical topic assignments [21]. Since its introduction, sLDA has become the foundation of various downstream topic models [32]. For example, Wang et al. extended the basic framework of multiclass sLDA to identify objective and subjective words in social media [40]. Ren et al. proposed a two-stage spectral method to recover the parameters and solve the local minimum defect of sLDA [31].

The upstream models assume topics are generated by metadata. Ramage et al. [29] proposed a model called Labeled LDA (L-LDA) for multi-labeled documents in which each document was correlated with several labels (e.g. books, science, grammar, and philosophy). L-LDA associates each word with a label and modifies LDA [2] by constraining a one-to-one correspondence between latent topics and users labels. To model the multi-labeled data, researchers have proposed methods such as partially labeled LDA (PLDA) [30] to understand the dependencies among the labels by projecting them onto a lower-dimensional latent space. They avoid the need to assume each label only relates to a single topic. Their experiments show that the allocation of multiple labels to one document improves model performance. Chauhan et al. proposed a labeled LDA model by considering additional constraints such as radius and time [5]. The proposed model can effectively predict the likely visiting places of users using their past tweets.

Although the upstream models perform well, they cannot be directly applied to our problem because these models are constructed based on the premise that each topic is related to only one label. This assumption is too restrictive and unreasonable to analyze the internal motivations of topic selection behavior, because behaviors are usually steered by the mixed effect of multiple intrinsic factors. In this paper, we develop a nonparametric hierarchical Bayesian topic model to identify the impact of intrinsic factors on users topic preference. We define the label-level (factor-level) random measures as draws from a Dirichlet Process in which the base measure is discrete instead of continues. The discrete attribute of

the Dirichlet Process allows topics to be shared across multiple intrinsic factors. Details of the proposed NHBT model are discussed next.

3. The proposed NHBT model

Suppose there are D users in a social media platform discussing various topics. Each user corresponds to a document which is the collection of the contents generated (posted, reposted, reviewed) by the user. The D documents contain V unique words or phrases. There are in all L intrinsic factors (e.g. age, gender, personality traits). The purpose of the NHBT model is to investigate how intrinsic factors affect topic selection. We first review the Dirichlet process (DP) and hierarchical Dirichlet process (HDP) which are the theoretical foundation of our new model. Then we detail the proposed NHBT model construction and inference procedure.

3.1. DP and HDP

Dirichlet Process (DP) is in essence a distribution over distributions. A Dirichlet process, denoted by $DP(\gamma, H)$, can be described by two parameters: a scaling (or concentration) parameter, $\gamma > 0$, and a base probability measure H over a space Ω. A draw from a Dirichlet process, $G_0 \sim DP(\gamma, H)$, is a distribution over Ω, such that, for any finite measurable partition $(A_1, A_2, \cdots, A_n)$ of Ω, the random vector $(G_0(A_1), G_0(A_2), \cdots, G_0(A_n))$ is distributed as

$$(G_0(A_1), G_0(A_2), \cdots, G_0(A_n)) \sim Dir(\gamma H(A_1), \gamma H(A_2), \cdots, \gamma H(A_n)) \tag{1}$$

The equation shows that the mean of draws from $DP(\gamma, H)$ will center around $(H(A_1), H(A_2), \cdots, H(A_n))$ and the concentration parameter determines the average deviation of samples from the base measure. From the perspective of topic model, Ω can be regarded as the set of all possible topics in documents and G_0 is the probability distribution on the topics. If we consider the influence of intrinsic factors on topic preferences, $(A_1, A_2, \cdots, A_n)$ can be regarded as the partition of topics resulted from the intrinsic factors. The distributions of topic partitions on the intrinsic factors can be obtained with Eq. (1).

The existence of DP was established by Ferguson [9]. Two concepts can help us understand the DP: one is the Chinese restaurant process and the other is the stick-breaking process (SBP). These two perspectives on Dirichlet process play a vital role in developing computational methods for Dirichlet process.

The Chinese restaurant metaphor is often used to explain the clustering effect of the Dirichlet process. It is a process of assigning restaurant tables to new customers, who enter a restaurant where an infinite number of tables exist. Suppose customers arrive one after another and sit down at tables chosen by customers according to certain process: (1) The first customer always chooses the first table. (2) The ith customer chooses a new unoccupied table with probability $\frac{\gamma}{i-1+\gamma}$, and an occupied table with probability $\frac{c}{i-1+\gamma}$,where c is the number customers sitting at that table.

More formally, we can denote the probability the ith customer will choose the kth table as

$$p(z_i = k) = \begin{cases} \frac{c_k}{\sum_t c_k + \gamma}, & for\ an\ occupied\ table \\ \frac{\gamma}{\sum_t c_k + \gamma}, & for\ an\ unoccupied\ table \end{cases} \tag{2}$$

where c_k refers to the number of customers sitting at table k.

The stick-breaking process (SBP) proposed by Sethuraman [35] provides an explicit way to construct a Dirichlet process. The property that measures drawn from a DP are discrete with probability 1, is made explicit in the SBP. Sethuraman [35] pointed out that a random sample G_0 drawn from $DP(\gamma, H)$ could be defined as

$$G_0 = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \tag{3}$$

where δ_{ϕ_k} is a probability measure concentrated at ϕ_k which are independent and identically distributed draws from the base measure H

$$\phi_k|_{k=1}^{\infty} \sim H \tag{4}$$

Parameter π_k is defined based on independent and identically distributed draws from $beta(1, \gamma)$

$$\pi_k = \pi_k' \prod_{l=1}^{k-1} \left(1 - \pi_l'\right)$$

$$\pi_k'|_{k=1}^{\infty} \sim beta(1, \gamma) \tag{5}$$

Note that G_0 in Eq. (3) is discrete with probability 1. Obviously, a draw from G_0 can be regarded as a distribution over infinite discrete "atoms". Each atom corresponds to a weight π_k and has a value ϕ_k drawn from the base measure H. The sequence $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ satisfies $\sum_{k=1}^{\infty} \pi_k = 1$, thus we can regard $\boldsymbol{\pi}$ as a random probability distribution on the positive integers, and usually represented as $\boldsymbol{\pi} \sim GEM(\gamma)$ [37].

In Bayesian formalism, a recurring theme is the need to partition the samples into groups, and yet allow the groups to remain linked to share statistical strength. However, the DP cannot fulfill the task. Hierarchical Dirichlet process, proposed

Table 1
Notations used in NHBT model.

Symbol	Description
D	Total number of documents
L	Total number of labels
V	Total number of unique words or phrases
γ	The concentration parameter for the first-level Dirichlet process
H	The base measure
α_0	The concentration parameter for the second-level Dirichlet process
G_0	The global random measures over topics drawn from H
α_l	The concentration parameter for the third-level Dirichlet process
G_l^L	The random measure over topics for label l drawn from G_0
G_d^D	The random measure over topics for document d
φ	Dirichlet prior for the topic-word distribution
η	Dirichlet prior for the document-label distribution
θ_d	Mixing proportion of labels for document d
F	A multinomial distribution that is conjugate to the base measure H
ϕ_k^G	The atom of G_0 drawn from H
β_k	The weight of atom k in G_0
ϕ_{lt}^L	The atom of G_l^L drawn from G_0
β_{lt}^L	The weight of atom t in G_l^L
ϕ_{dr}^D	The atom of G_d^D drawn from mixing G_l^L's
β_{dr}^D	The weight of atom r in G_d^D
k_{dn}^D	Index of table assigned to token x_{dn}
k_{dj}^L	Index of label-level table assigned to table j of document d
$k_{i=(l,t)}^G$	Index of global-level table assigned to table t of label l
$N_{k,w}$	The number of times word type w in the vocabulary is assigned to topic k and $N_{k,\cdot}$ is the margin count
$N_{d,j}$	The number of words that belong to table j in document d
$M_{d,l}$	The number of tables that belong to label l in document d
$M_{\cdot,l,t}$	The number of document-level tables assigned to table t of label l
M_k	The number of label-level tables assigned to global-level table k

by Teh et al. [37], is a hierarchical Bayesian model that can be successfully used to tackle group data.Hierarchical models share the parameters across groups and the randomness of parameters induces dependencies among the groups.HDP introduces a set of random measures G_d, one for each group d. The core of HDP is that the base measure G_0 for the group-level Dirichlet processes G_d's is itself distributed according to a Dirichlet process $G_0 \sim DP(\gamma, H)$. Since such base measure is discrete, the group-level Dirichlet processes have to share atoms. More formally, we denote the HDP using a stick-breaking representation.

$$\begin{aligned} G_0 &\sim DP(\gamma, H) \qquad G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\ G_d &\sim DP(\alpha_0, G_0) \qquad G_d = \sum_{s=1}^{\infty} \pi_{ds} \delta_{\phi_{ds}} \end{aligned} \tag{6}$$

where $\phi_k \sim H$, $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \sim GEM(\gamma)$, $\phi_{ds} \sim G_0$, and $\boldsymbol{\pi} = (\pi_{ds})_{s=1}^{\infty} \sim GEM(\alpha_0)$.

3.2. The nonparametric hierarchical Bayesian topic (NHBT) model

To model the generative relationships from the intrinsic factors to topic preferences, we suppose the global random measure G_0 is sampled from standard Dirichlet Process $DP(\gamma, H)$. Each document d corresponds to a specific random measure G_d^D over topics, and each intrinsic factor has a specific random measure G_l^L over topics. Table 1 provides a summary of notations used in the proposed NHBT model.

The NHBT model is a hierarchical Bayesian model that has three-level random measures. The first level (*global level*) defines the global random measure G_0, which is a draw from a standard Dirichlet process $DP(\gamma, H)$. The base distribution H is assumed to be a symmetric Dirichlet distribution over a fixed vocabulary dimension. In the second level (*factor level*), we define a set of DP distributed random measures $\left(G_0^L, G_1^L, \cdots, G_L^L\right)$ over L possible intrinsic factors with a base distribution G_0 according to $G_l^L | \alpha_0, G_0 \sim DP(\alpha_0, G_0)$ where G_0 is captured in the global level. The concentration parameter α_0 controls the variability of G_l^L. We define an infinite topic space for each intrinsic factor by setting one random measure per factor. Since the random measure G_0 is discrete, our model makes sure that a single topic can be shared across multiple labels. We can thus assume that one topic can be generated by multiple intrinsic factors, a scenario not addressed by traditional Bayesian topic models for metadata. In the third level (*document level*), we define a set of DP distributed random measures $\left(G_0^D, G_1^D, \cdots, G_D^D\right)$ for the D documents, which regard a mixture of random measures as a base distribution of the Dirichlet process.

We now detail the NHBT model. For the *global level*, the general random measure G_0 follows the standard Dirichlet process, which is given by Eq. (7):

$$\begin{aligned} &G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k^G} \\ &\phi_k^G|_{k=1}^{\infty} \sim H \\ &\beta_k|_{k=1}^{\infty} \sim beta(1,\gamma) \end{aligned} \tag{7}$$

The distribution G_0 defines the global random measure over topics drawn from the base probability measure H. Since each atom ϕ_k^G corresponds to a topic, G_0 determines all the possible topics in documents.

For the *factor level*, the set of random measures for L factors is also drawn from standard Dirichlet process in which the base measure is G_0.

$$\begin{aligned} &G_l^L = \sum_{t=1}^{\infty} \beta_{lt}^L \delta_{\phi_{lt}^L} \\ &\phi_{lt}^L \sim G_0 \\ &\beta_{lt}^L = \tau_{lt}^L \prod_{e=1}^{t-1} \left(1-\tau_{le}^L\right) \\ &\tau_{lt}^L \sim beta(1,\alpha_0) \end{aligned} \tag{8}$$

As shown in Eq. (8), the *factor level* draws atoms ϕ_{lt}^L from a Dirichlet process in which G_0 is the base measure. Because samples drawn from Dirichlet process are discrete with probability 1 and G_0 is sampled from a standard Dirichlet process, G_0 is discrete in our model. If we draw topic atoms independently for $G_l^L, l = 1, 2, \cdots, L$ from G_0, some atoms would be shared among the factor-level random measures. The restriction in traditional Bayesian topic model that one topic is related to a single intrinsic factor can be released.

For the *document level*, the stick breaking construction is given by the following conditional distributions:

$$\begin{aligned} &\theta_d \sim Dir(\rho_d \eta) \\ &\varpi_{dr} \sim beta(1,\alpha_1) \\ &\beta_{dr}^D = \varpi_{dr} \prod_{p=1}^{r-1} \left(1-\varpi_{dp}\right) \\ &h_{dr} \sim Mult(\theta_d) \\ &\phi_{dr}^D \sim G_{h_{dr}}^L \\ &G_d^D = \sum_{r=1}^{\infty} \beta_{dr}^D \delta_{\phi_{dr}^D} \end{aligned} \tag{9}$$

From Eq. (9) we can see that each atom ϕ_{dr}^D in document D is determined by a random measure $G_{h_{dr}}^L$ in the *factor level*. It indicates the assumption of multiple-to-multiple relationship used in the proposed model. That is, the atoms of a document are generated by atoms of multiple factors in the *factor level* and an atom in the factor level has influence on the topics in multiple documents.

After obtaining random measures G_d^D 's for each documents, we can apply it to generate the words in the documents. We first draw an indicator variable ψ_{dn} from G_d^D, and then generate the word x_{dn} with a distribution F based on ψ_{dn}. Distribution F is usually assumed as a multinomial distribution and is conjugate to the base measure H. This strategy makes the computation easy since we can integrate out ψ_{dn}.

$$\begin{aligned} &\psi_{dn}|G_d^D \sim G_d^D \\ &x_{dn}|\psi_{dn} \sim F(\psi_{dn}) \end{aligned} \tag{10}$$

Different with the standard Dirichlet process, our model adopts the Dirichlet Process with mixed random measures to assign "atoms" to each "stick" [17]. That is, parameter h_{dr} is used as an indicator to G_d^D where atom ϕ_{dr}^D is drawn. Each ψ_{dn} corresponds to one ϕ_{dr}^D (i.e., $\psi_{dn} = \phi_{dr}^D$), each ϕ_{dr}^D is associated with one ϕ_t^L, and each ϕ_t^L is correlated with one ϕ_k^G.

Fig. 1 illustrates a sharing structure among the three-level random measures. The generation process of the variables used in the proposed NHBT model is shown in Fig. 2.

3.3. Inference procedure of NHBT model

In this section, we design an inference procedure to solve the proposed NHBT model based on Gibbs sampling. The state space of our chain includes the topic $\boldsymbol{z} = \{z_{d,n}\}$ and intrinsic factor $\boldsymbol{l} = \{l_{d,n}\}$, assigned to all words x_{dn}. Since we have

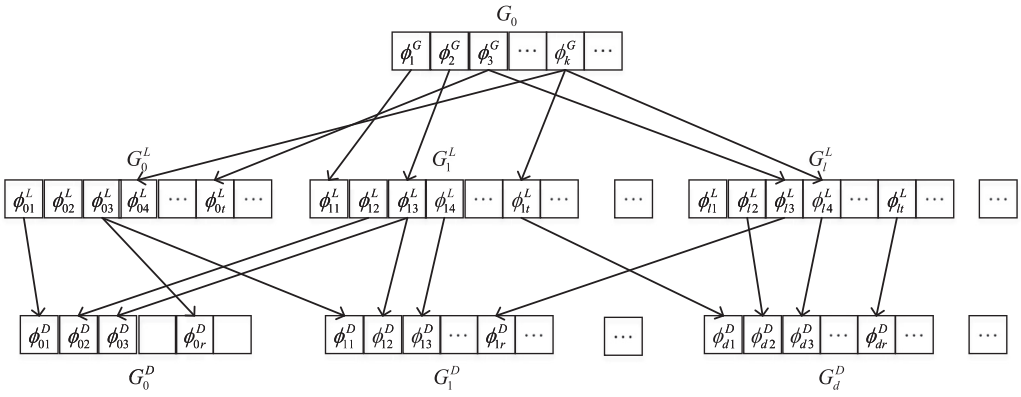

Fig. 1. Sharing structure among the three-level random measures.

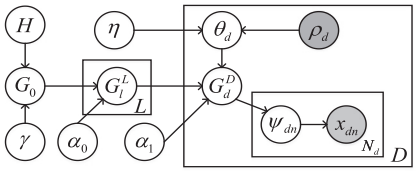

Fig. 2. Generation process of variables.

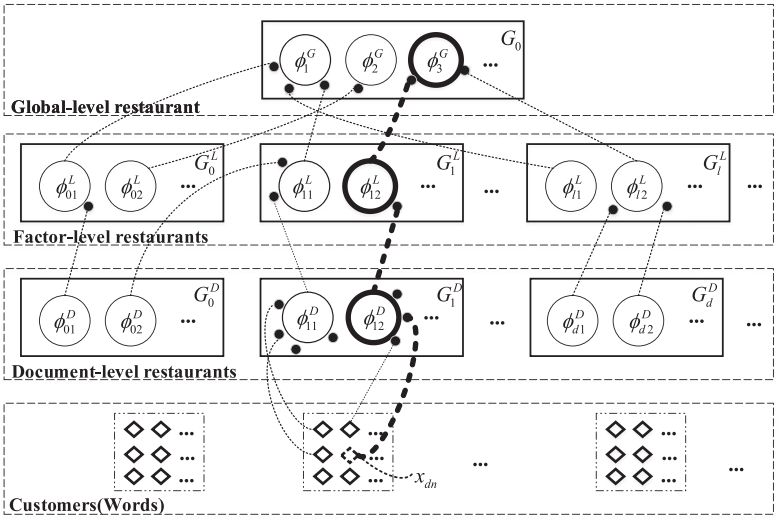

Fig. 3. Illustration of topic assignments in our inference algorithm.

three-level random measures, to obtain $z_{d,n}$ and $l_{d,n}$, we must know the path assigned to each word x_{dn} in the three-level hierarchical structure that includes table assignments k^D_{dn} for document level, k^L_{dj} for factor level, and k^G_i for global level. The sampling scheme is illustrated in Fig. 3. The core of our inference procedure is to use the Chinese restaurant process to integrate out the three level random measures, i.e., G_0, G^L_l, and G^D_d, and the other latent variables, e.g. θ_d and ϕ_k.

In Fig. 3, each bold solid rectangle can be considered as a restaurant and each circle in the rectangles can be considered as a table. To integrate the variables within the three-level framework, we sample a table index in each level for word x_{dn} based on Chinese restaurant process. In document level, we first sample a table index k^D_{dn} for word x_{dn}. The word x_{dn} can belong to either an occupied table j or a new unoccupied one j^{new}. If a new document-level table is sampled, a new table index $k^L_{dj^{new}}$ in factor-level restaurant will be created. The table index $k^L_{dj^{new}}$ contains two elements: intrinsic factor index and table index given the intrinsic factor index. Therefore, we need sample new factor and table indices in factor-level restaurant for the j^{new} table in doucument-level. Similarly, the new table index $k^L_{dj^{new}}$ in factor level can also be either an existing factor-level table i or a new one i^{new}. If i^{new} is sampled, we assign an existing topic k or a new one k^{new} in global

level. The bold dashed lines show a sampling path for the word x_{dn}. In this example, we sample 2 for $k^D_{d,n}$, (1,2) for $k^L_{d,j}$, and 3 for k^G_i. Therefore, the intrinsic factor and topic indices of this word are $(l_{dn}, z_{dn}) = (1, 3)$.

To estimate the generative relationships between intrinsic factors and topic selection behaviors using the above process, we sample the three table indices $k^D_{d,n}$, $k^L_{d,j}$, and k^G_i. Namely, we need to know the topic of this word and which factor trigger this word. Details of parameter calculation within Bayesian framework are presented next.

3.3.1. Conditional density $f_k^{-d,n}$

Before we provide the probability equations, we establish the conditional density $f_k^{-d,n}$ of x_{dn} under topic k for all data items except x_{dn}.

$$\begin{aligned} f_k^{-d,n}(x_{dn} = w) &= \frac{\int_{\phi^G_k} p\left(x_{dn}|\phi^G_k\right) \prod_{d'n' \neq dn, z_{d'n'=k}} p\left(x_{d'n'}|\phi^G_k\right) p\left(\phi^G_k|\varphi\right) d\phi^G_k}{\int_{\phi^G_k} \prod_{d'n' \neq dn, z_{d'n'=k}} p\left(x_{d'n'}|\phi^G_k\right) p\left(\phi^G_k|\varphi\right) d\phi^G_k} \\ &= \begin{cases} \frac{N_{k,w}+\varphi}{N_{k,\bullet}+V\varphi}, & if\ k\ exists \\ \frac{1}{V} & if\ k\ is\ new \end{cases} \end{aligned} \tag{11}$$

where $N_{k,\,w}$ represents the number of word type w assigned to topic k, $N_{k,\cdot}$ is the total number of words that belong to topic k in D documents, and superscript $-d,n$ denotes the same count excluding x_{dn}.

The conditional density $f_k^{-d,n}$ can be considered as the likelihood function to sample $k^D_{d,n}$, $k^L_{d,j}$ and k^G_i, and can be calculated based on statistic counts.

3.3.2. Sampling $k^D_{d,n}$

The probability of assigning a document-level table j to word x_{dn} is

$$\begin{aligned} p\left(k^D_{d,n} = j|\boldsymbol{k}^D_{-(d,n)}, x_{d,n}, \boldsymbol{x}_{-(d,n)}, rest\right) &\propto p\left(k^D_{d,n} = j|\boldsymbol{k}^D_{-(d,n)}\right) p\left(x_{d,n}|k^D_{d,n} = j, \boldsymbol{k}^D_{-(d,n)}, \boldsymbol{x}_{-(d,n)}, rest\right) \\ &= \begin{cases} \frac{N_{d,j}}{N_{d,\bullet}+\alpha_1} f_{k^G_{k^L_{d,j}}}\left(x_{d,n}\right), & if\ j\ exists \\ \frac{\alpha_1}{N_{d,\bullet}+\alpha_1} \Gamma_0\left(x_{d,n}\right), \end{cases} \end{aligned} \tag{12}$$

where $N_{d,\,j}$ denotes the number of words that belong to table j, $N_{d,\cdot}$ is number of words in document d, $N_{d,\bullet} = \sum_j N_{d,j}$, α_1 is the concentration parameter for the third-level Dirichlet process, T_l denotes the number of tables in the label-level restaurant l, and $\Gamma_0(x_{d,n}) = \sum_{l=1}^{L} \frac{M_{d,l,\bullet}+\eta}{M_{d,\bullet,\bullet}+L_d\eta}[\sum_{t=1}^{T_l} \frac{M_{\bullet,l,t}}{M_{\bullet,l,\bullet}+\alpha_0} f_{k^G_{t,t}}(x_{d,n}) + \frac{\alpha_0}{M_{\bullet,l,\bullet}+\alpha_0}\Gamma_1(x_{d,n})]$.

3.3.3. Sampling $k^L_{d,j}$

When a new document-level table j^{new} is sampled, we need to associate it with a table in the label-level restaurant. More specifically, we sample a combination of label l and table t for that new document-level table. We use index i to denote the combination, $i = (l, t)$. The probability of assigning the table j^{new} to a label-level combination i is

$$\begin{aligned} & p\left(k^L_{d,j^{new}} = (l,t) = i|\boldsymbol{k}^L_{-(d,j^{new})}, x_{d,n}, \boldsymbol{x}_{-(d,n)}, rest\right) \\ & \propto p\left(k^L_{d,j^{new}} = (l,t) = i|\boldsymbol{k}^L_{-(d,j^{new})}\right) p\left(x_{d,n}|k^L_{d,j^{new}} = (l,t) = i, \boldsymbol{k}^L_{-(d,j^{new})}, \boldsymbol{x}_{-(d,n)}, rest\right) \\ &= \begin{cases} \frac{M_{d,l,\bullet}+\eta}{M_{d,\bullet,\bullet}+L_d\eta} \frac{M_{\bullet,l,t}}{M_{\bullet,l,\bullet}+\alpha_0} f_{k^G_i}\left(x_{d,n}\right) \\ \frac{M_{d,l,\bullet}+\eta}{M_{d,\bullet,\bullet}+L_d\eta} \frac{\alpha_0}{M_{\bullet,l,\bullet}+\alpha_0} \Gamma_1\left(x_{d,n}\right) \end{cases} \end{aligned} \tag{13}$$

where $M_{d,\,l,\,\cdot}$ denotes the number of tables that belong to label l in document d, $M_{d,\,\cdot,\,\cdot}$ is the total number of tables in document d, L_d refers to the number of labels in document d, $M_{\cdot,\,l,\,t}$ represents the number of customers that belong to label l and table t, $M_{\cdot,\,l,\,\cdot}$ is the total number of customers that belong to label l, α_0 refers to the concentration parameter for the factor-level Dirichlet process, and $\Gamma_1\left(x_{d,n}\right) = \sum_{k=1}^{K} \frac{M_k}{M_\bullet+\gamma} f_k\left(x_{d,n}\right) + \frac{\gamma}{M_\bullet+\gamma} f_{k^{new}}\left(x_{d,n}\right)$.

Input: D documents and their intrinsic factors
Output: Generative relationship from intrinsic factors to topics

Step 1 Model construction

Step 1.1: Construct global level random measure G_0 by Equation (7);
Step 1.2: Construct factor level random measures $(G_0^L, G_1^L, \ldots, G_L^L)$ by Equation (8);
Step 1.3: Construct document level random measures $(G_0^D, G_1^D, \ldots, G_D^D)$ by Equation (9);
Step 1.4: Generate topics and words based on document level measures by Equation (10);

Step 2 Model inference

Step 2.1: Establish the conditional density $f_k^{-d,n}$ by Equation (11);
Step 2.2: Sampling $k_{d,n}^D$ by Equation (12);
Step 2.3: Sampling $k_{d,j}^L$ by Equation (13);
Step 2.4: Sampling k_i^G by Equation (14).

Fig. 4. Model framework.

3.3.4. Sampling k_i^G

When a new factor-level table i^{new} is sampled, we need to associate it with a topic in the global-level restaurant. The sampling formula is as follows.

$$p\left(k_i^G = k|\boldsymbol{k}_{-i}^G, x_{d,n}, \boldsymbol{x}_{-(d,n)}, rest\right) \propto p\left(k_i^G = k|\boldsymbol{k}_{-i}^G\right) p\left(x_{d,n}|k_i^G = k, \boldsymbol{k}_{-i}^G, \boldsymbol{x}_{-(d,n)}, rest\right)$$

$$= \begin{cases} \dfrac{M_k}{M_\bullet + \gamma} f_k\left(x_{d,n}\right) \\ \dfrac{\gamma}{M_\bullet + \gamma} f_{k^{new}}\left(x_{d,n}\right) \end{cases} \tag{14}$$

where M_k denotes the number of customers in table k (i.e., the total number of tables that belong to topic k in label-level restaurants), $M_\cdot$ is the total number of tables in label-level restaurants, and γ is the concentration parameter for the first-level Dirichlet process.

3.4. Model framework

With the Equations of the NHBT model and inference procedure, we can follow the steps in Fig. 4 to analyze the generative relationship from intrinsic factors to topics.

4. Experiments and results

In this section, we test the proposed NHBT model based on the real data from myPersonality, which is a popular Facebook application allowing users to take a set of psychological questionnaires and share results with their friends [18]. To preprocess myPersonality data, we remove the stop words, punctuations, numbers, URLs, and characters not in Latin from the corpus. We conduct standard stemming operation to alleviate data sparseness problem. We also consider bigrams which are constructed by two words.

We obtain 7508 users satisfying the following three conditions: (1) Users ages are 20 or 21. Because young people are the main group in Facebook, we regard age as a control variable. (2) Users have published at least 1000 words in their Facebook status updates. We combine all status updates of a user into a single document, as the length of a Facebook status update is usually short which often hurts the performance of topic models [19]. (3) Users have taken the five major personality questionnaires based on personality test.

The questionnaires employed in myPersonality classify personality into five dimensions:

a) openness (e.g., adventurous, curious, imaginative),
b) conscientiousness (e.g., reliable, self-disciplined, ambitious),
c) agreeableness (e.g., sympathetic, cooperative, trusting, kind),
d) extraversion (e.g., optimistic, active, talkative), and
e) neuroticism (e.g., sadness, embarrass, anxious).

Because the original personality traits are continuous, we employ the following method to discretize them. First we calculate mean and standard deviation for each dimension of the five personality dimensions. Then we use these two values to discretize each dimension into three levels (low, middle and high).

Besides the five dimensions of the personality model, we also incorporate gender as an intrinsic factor because it is a well-documented indicator for user online behavior [39,42]. Therefore, each user in our experiment is characterized by these six intrinsic factors. The 17 labels contained in the six intrinsic factors are shown in Table 2.

Note that the proposed NHBT model is indeed a language model in which topics are extracted based on the word frequencies in the dataset. Therefore, some words (e.g. “love”) that are frequently used may appear in top positions among

Table 2
The labels of users.

Intrinsic factors	Labels and values			
Gender	Label	Male	Female	
Openness	Label	HighOp	MiddleOp	LowOp
	Value	[4.56, 5]	[3.22, 4.56)	[1, 3.22)
Conscientiousness	Label	HighCo	MiddleCo	LowCo
	Value	[4.10, 5)	[2.63, 4.10)	[1, 2.63)
Agreeableness	Label	HighAg	MiddleAg	LowAg
	Value	[4.41, 5)	[2.80, 4.41)	[1, 2.80)
Extraversion	Label	HighEx	MiddleEx	LowEx
	Value	[4.26, 5)	[2.85, 4.26)	[1, 2.85)
Neuroticism	Label	HighNe	MiddleNe	LowNe
	Value	[3.61, 5)	[2.00, 3.61)	[1, 2.0)

Table 3
Influence of gender on topic preference.

Male			Female		
Topic 1 (politics)	Topic 2 (military)	Topic 3 (football)	Topic 4 (family)	Topic 5 (daily routines)	Topic 6 (study)
obama	alpha_novemb	liverpool	babi_girl	dayi	roommat
republican	marin	england	hubbi	youu	midterm
presid	militari	world_cup	husband	todayi	professor
liberti	yanke_oscar	chelsea	babi_shower	funn	emili
liber	delta_papa	Fifa	pregnant	night	scholarship
elect	warship	spain	babi_boi	outt	calculu
conserv	papa_yanke	roonei	famili	gunna	bio
democrat	uniform_romeo	barca	due_date	bedd	research_paper
congress	tango_oscar	soccer	man	tonight	chem
justic	tango_alpha	brazil	gna	work	textbook

all topics. To make characteristic words of specific topics more apparent, we calculate the term score of each word in each topic based on Eq. (15).

$$\begin{aligned} terms\_core_{k,v} &= \hat{\phi}_{k,v} log\left(\frac{\hat{\phi}_{k,v}}{\left(\prod_{j=1}^{K}\hat{\phi}_{j,v}\right)^{\frac{1}{K}}}\right) = \hat{\phi}_{k,v}\left[log\left(\hat{\phi}_{k,v}\right) - log\left(\left(\prod_{j=1}^{K}\hat{\phi}_{j,v}\right)^{\frac{1}{K}}\right)\right] \\ &= \hat{\phi}_{k,v}\left[log\left(\hat{\phi}_{k,v}\right) - \frac{1}{K}log\left(\prod_{j=1}^{K}\hat{\phi}_{j,v}\right)\right] \\ &= \hat{\phi}_{k,v}\left[log\left(\hat{\phi}_{k,v}\right) - \frac{1}{K}\sum_{j=1}^{K}log\left(\hat{\phi}_{j,v}\right)\right] \end{aligned} \tag{15}$$

where $\hat{\phi}_{k,v}$ denotes the probability that word v belongs to topic k. Because frequent words usually appear in the top positions among almost all topics, some words which are not frequent in all topics but important for specific topics would be excluded from the discovered topics if we maintain these frequent words. It would result in negative influence on model performance. Therefore, by assigning lower scores to the words that cross frequently multiple topics and higher scores to the words that occur exceptionally often in specific topics [15,36], Eq. (15) can help us obtain more accurate topics and enhance our model performance.

4.1. Topics generated from gender

This section examines the influence of gender on topic preferences. The typical topics extracted by the NHBT model are shown in Table 3. Table 3 shows that NHBT model can extract topics that associated with gender. The listed six topics can be tagged as politics (Topic 1), military (Topic 2), football (Topic 3), family (Topic 4), daily routines (Topic 5), and study (Topic 6), respectively. Topic preferences are obviously different between male and female users. Table 3 shows the men tend to focus on politics, military, and football, while women prefer topics on family, daily routines, and study.

Table 4
Example of topics extracted by NHBT model about openness and conscientiousness.

LowOp	LowOp	LowOp	HighOp	HighOp	LowCo	LowCo	HighCo	HighCo	HighCo
Topic 7	Topic 8	Topic 4	Topic 9	Topic10	Topic11	Topic12	Topic 6	Topic 1	Topic13
basketbal	snowboard	babi_girl	pokemon	Reed	wwe	princess_gan	roommat	obama	drill
coach	canada	Hubbi	xbox	march_band	wrestl	commerc	midterm	republican	brick
yanke	skate	Husband	metal	Piano	smackdown	rio	professor	presid	hammer
hockei	tierd	babi_shower	sword	band_camp	wrestlemania	ddr	emili	liberti	peg
playoff	whistler	Pregnant	japanes	Audit	nick	packet	scholarship	liber	mallet
nfl	happi_fridai	babi_boi	bastard	Trumpet	bridge	party	calculu	elect	shingl
championship	whoo	family	video_game	Clarinet	arm	school_start	bio	conserv	massag
defens	ski	due_date	halo	Orchestra	crazi	hot_chocol	research_paper	democrat	cement
basebal	snow	Man	alien	Rehears	lock	abbi	chem	congress	massag_therapi
football_game	run	Gna	weapon	jazz_band	fuckin	tango	textbook	justic	chicken_cook

Table 5
Example of topics extracted by NHBT model about extraversion, agreeableness, and neuroticism.

LowEx	LowEx	HighEx	HighEx	LowAg	HighAg	HighAg	LowNe	LowNe	HighNe	HighNe
Topic 9	Topic14	Topic15	Topic12	Topic16	Topic17	Topic18	Topic19	Topic20	Topic21	Topic22
pokemon	chem	revisit	princess_gan	Omfg	video_chat	allah	ahaha	christ	gaga	cosplai
xbox	bio	england	commerc	cigarett	web	muslim	aha	prais	fame	miku
metal	orgo	hotel	rio	bleh	hei	moham	yeee	glori	rihanna	sim
sword	chemistri	tourist	ddr	fuck_hate	chatroom	arab	hah	worship	luv	japanes
japanes	calc	attract	packet	god_damn	imma	egypt	ahahaha	god_bless	concert	kitten
bastard	lab_report	pub	party	watch_crimin	vid_chat	india	hahah	merci	beyonc	kenni
video_game	boiler	travel	school_start	head_hurt	tinychat	eid	ahah	psalm	michael_jackson	xdd
halo	chem_test	fanci	hot_chocol	dick	luv	ramadan	smh	proverb	taylor_swift	mah
alien	biologi	ticket	abbi	bullshit	tinychat_time	egyptian	hella	presenc	song	shaun
weapon	bry	bus	tango	whore	live_video	sheesh	yeeee	spiritu	cat_steven	kitti

4.2. Topics generated from personality traits

This section explores the internal motivation of topic preference from the perspective of personality traits. Eq. (16) gives the relationships between intrinsic factors and topic preferences.

$$p(l|z) \propto p(z|l)p(l) \tag{16}$$

From Eq. (16), we can obtain the weight distribution of intrinsic factors $l, l = 1, 2, \cdots, L,$ on topic z. Large weights indicate the corresponding factors are more relative to the topic. We assume $p(l)$ is uniform [41].

The existing literatures show that, compared with the relationships between user behavior and lower or higher personality, the relationships between middle personality and user behaviors are usually unreliable and inconvincible. Therefore, to be consistent with the existing literatures, we only report the topic preferences related to lower or higher personality [34]. Tables 4 and 5 provide the topics preferred by users who have high or low levels in the five personality dimensions (see Table 2). Through the proposed NHBT model, we uncover the personality-related topics that cannot be found by the closed-vocabulary framework. For example, users with high openness prefer topics in computer games (Topic 9) and music band (Topic 10). Introverts prefer Chemical Biology (Topic 14) besides computer games. Users low in agreeableness are more likely to publish swear words (Topic 16) and users high in agreeableness like to use video chat (Topic 17) to communicate with others. An interesting finding is that neurotic users like to play Cosplay (Topic 22), an exciting game. Besides the interesting topics implied in the open-vocabulary, our model also has the ability to discover the explicit topics embedded in the LIWC software, e.g. family (Topic 4) and religion (Topic 20). Our model shows that people with high conscientiousness tend to post more words about family and people with low neuroticism (i.e., emotion stability) are more likely to publish topics about religion.

All these findings are reasonable, meaningful, and close to users real life. As free-form languages are adopted by social network users, the dictionary in LIWC software cannot capture all contents user wrote and thus cannot accurately depict online users behavior. The proposed NHBT model is an open-vocabulary-driven method and can automatically extract these latent topics.

An assumption of the NHBT model is that topic preference is the external behavior impacted by multiple intrinsic factors. Our experiment shows that the intrinsic factors play distinct roles in topic selection. As we can see from Table 4 and 5, Topic 4 (family topic) is generated by female, LowOp and HighAg labels. Topic 6 (study topic) is generated by the female and HighCo labels, and Topic 1 (politics) is generated by the male, HighCo and LowOp labels. Similarly, both the HighOp and LowEx labels contribute to Topic 9 (computer games), and both the LowCo and HighEx labels result in Topic 12 (campus party).

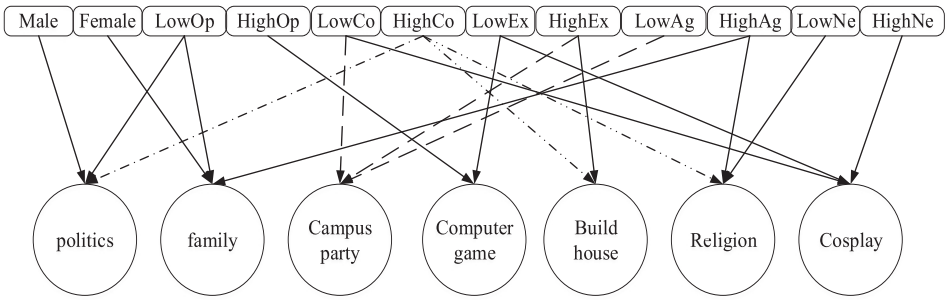

Fig. 5. Generative relationship from intrinsic factors to topics.

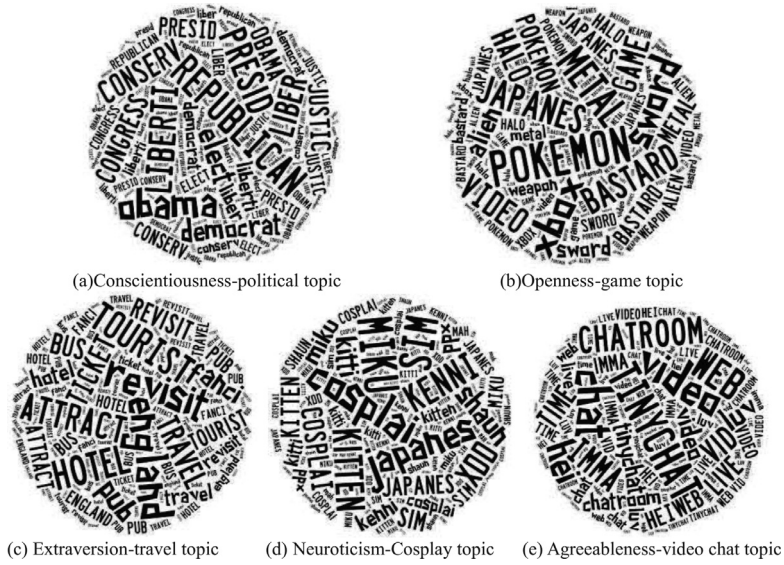

Fig. 6. Word clouds from personality traits to topics.

Table 6
Label distributions for topics.

Topics	Weight
Topic 1(politics)	HighCo(0.171), Male(0.130), LowOp(0.113)
Topic 4(family)	LowOp(0.189), Female (0.164), HighAg(0.147)
Topic 12(campus party)	HighEx(0.120), LowCo(0.115), LowAg(0.109)
Topic 13(build house)	HighCo(0.457), HighEx(0.105)
Topic 20(religion)	HighAg(0.120), LowNe(0.119), HighCo(0.115)
Topic 22(Cosplay)	HighNe(0.210), LowEx(0.206), LowCo(0.105)

Fig. 5 illustrates the generative relationship from intrinsic factors to social media topics. As shown in Fig. 5, the intrinsic factor HighCo can generate multiple topics about politics, build house, and religion. On the contrary, the topic on campus party is generated by multiple intrinsic factors such as LowCo, HighEx and LowAg. The proposed NHBT model can effectively capture the multiple-to-multiple generative relationships. Fig. 6 illustrates five examples of word clouds from personality traits to topic preferences.

Table 6 provides the weight distributions of the labels which play primary roles on generating Topic 1, 4, 11, 13, 20 and 22. The weight distributions of all the labels for Topic 1 and 4 are displayed in Fig. 7. Although all the intrinsic factors impact topic preference, the primary labels and their contributions are different for distinct topics. For example, for Topics 1 and 4, the primary intrinsic labels are (HighCo, Male, LowOp) and (LowOp, Female, HighAg), respectively. The contributions of the labels on generating topic preferences are quantified by weights (0.171, 0.130, 0.113) and (0.189, 0.164, 0.147), respectively. The relationships presented in Table 6 and Fig. 7 are partially consistent with the literature. For example, Schwartz et al. [34] maintain people high in agreeableness and conscientiousness prefer religion topics. Family topics are associated with females [25], people low in openness and high in agreeableness [44]. Compared with the existing literature, the proposed NHBT model obtains more detailed insights.

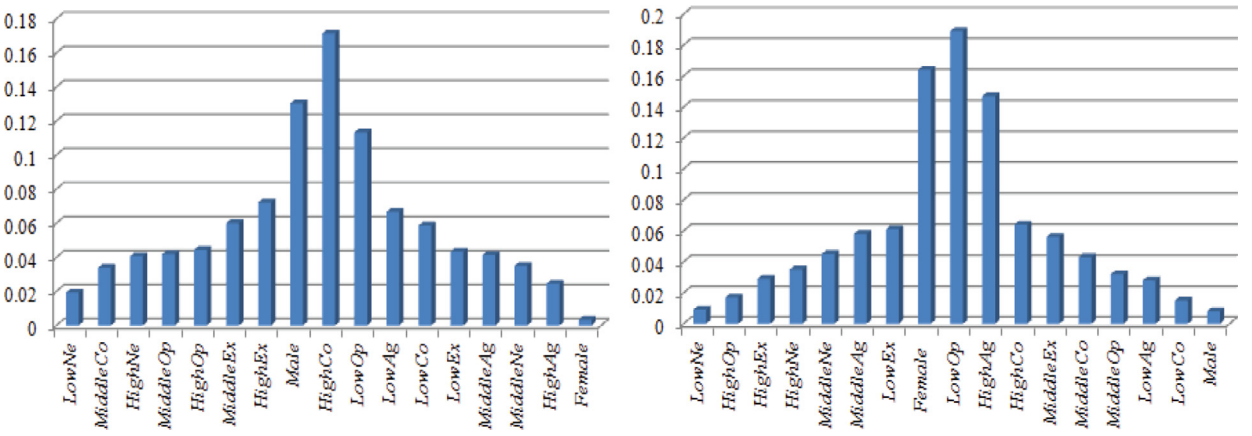

Fig. 7. The label distributions for Topic 1 and 4.

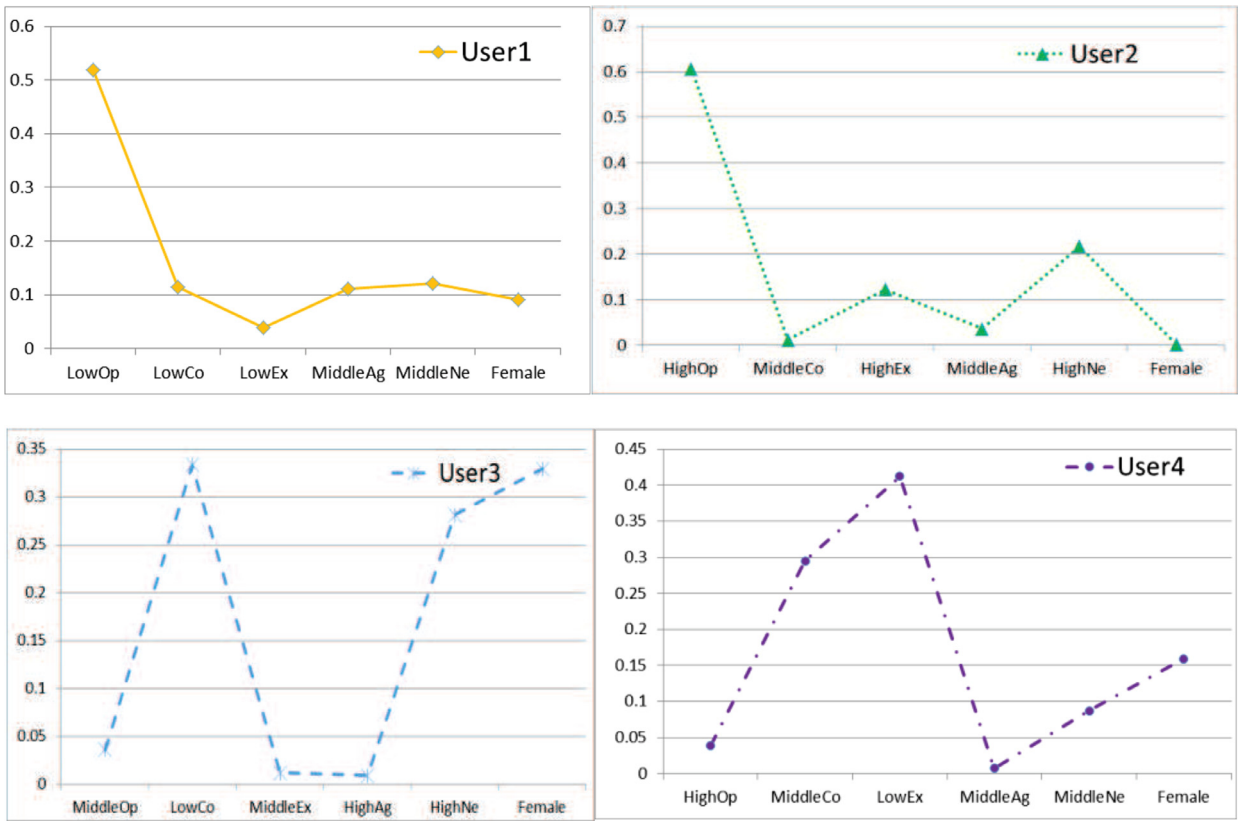

Fig. 8. Label distributions for individual users.

4.3. Label distributions for individual user

The intrinsic factors play different roles for topic selection, and their influences on individual users are also significantly different. Eq. (17) identifies the factor distribution for individual users who have exposed their personality traits in myPersonality.

$$p(l|d) = \frac{M_{d,l,\bullet}+\eta}{M_{d,\bullet,\bullet}+L_d\eta} \tag{17}$$

where d is the document generated by a user. Fig. 8 shows the weight distributions of the intrinsic factors for four individuals. With the proposed NHBT model, we can identify the important labels for individual users. For example, the openness

Table 7
Prediction accuracy comparison of NHBT, SVM and NB method.

	Gender	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
NHBT	0.8808	**0.6691**	**0.6232**	**0.6438**	**0.6125**	0.5925
SVM	0.8602	0.6498	0.5932	0.6385	0.5872	**0.6012**
NB	0.8695	0.6318	0.6005	0.6212	0.5726	0.5899

personality almost dominates topic preference behavior of User 1, while the behavior of User 3 is under the control of three intrinsic factors such as conscientiousness (LowCo), extraversion (HighNe) personalities and gender (Female). These results offer insights into individual behavior. Also, to predict a users subsequent behavior, we can use the dominant intrinsic factors instead of the factors with minor influence to enhance computational efficiency.

4.4. User label prediction

With the proposed NHBT model, we can obtain the generative relationships between intrinsic factors and topic selection behaviors. For new users who have posted contents but not exposed their personality traits, managers may be more interested in the internal factors behind the contents. The prediction of users inherent labels is an important issue in the field of behavior analysis [18,26]. The NHBT model provides foundation to extract the internal motivation behind user generated contents. The label distributions for new user generated contents can be estimated by Eq. (18).

$$p(l|d^*) \propto p(d^*|l) = \prod_{x \in x_{d^*}} p(x|l) = \prod_{x \in x_{d^*}} \sum_{z=1}^{L_z} p(x|z,l)p(z|l) \tag{18}$$

where d^* is the document for the new user and L_z denotes the number of topics associated with label l.

Our experiment compares the estimated values with the real values, calculate the estimation accuracy and compare the accuracy with baseline methods. We choose Support Vector Machine (SVM) and Naive Bayes (NB) as our baselines because both approaches are widely used in classification problem and have shown great performance. The SVM algorithm used in our experiment is from software Libsvm [4]. We use the embedded C-SVC algorithm with the radial basis function in our experiment. The NB algorithm is a straightforward method and we use the Naive Bayes package of JAVA [11] in our experiment. Both of the algorithms are implemented with the default parameters in the corresponding software to build the personality classifier. The comparison results presented in Table 7 show that the proposed NHBT model can predict intrinsic labels based on user generated contents. It outperforms SVM and NB approach except for neuroticism. Table 7 also shows that openness personality is the easiest to predict while predicting neuroticism personality is difficult. For users with high Neuroticism level, behavior dissimilation makes them act inconsistently in the online and offline environment. They apt to pretend their online behaviors [43] and use asynchronous communication to earn more time to decide what should say and how to express in the online environment [33]. Behavior dissimilation easily misguides us to exact confusing features from online social media data and thus negatively impact the prediction accuracy.

4.5. Predictive performance

Different with the models such as L-LDA and PLDA which assume one-to-one relationship between intrinsic factors and topics, the proposed NHBT model is constructed on the basis of multiple-to-multiple relationship. To test the effectiveness

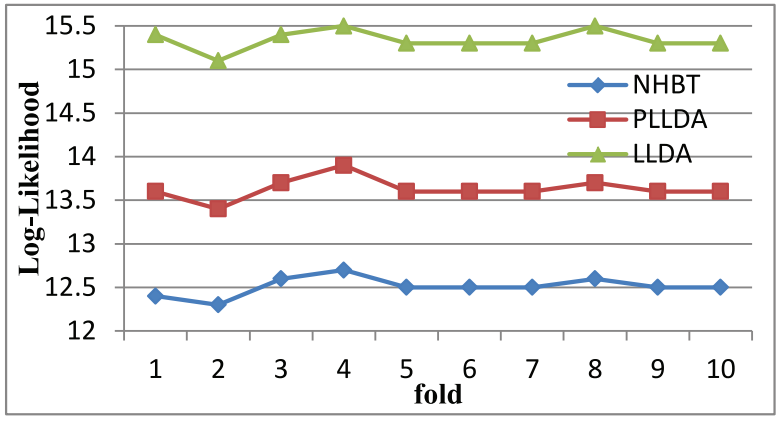

Fig. 9. Held-out Log-Likelihood results.

of our assumption, we employ the three models (i.e. NHBT, L-LDA and PLDA) to fit the real personality values and compare their performance by the held-out Log-Likelihood value. We use 90% of the data for training and the rest of 10% for testing. Fig. 9 shows that the performance of our model is significantly better than the performance of L-LDA and PLDA. It proves that the multiple-to-multiple assumption accurately reflects the relationship from the intrinsic factors to topic preferences.

5. Conclusions and future work

This paper proposes a nonparametric hierarchical Bayesian topic (NHBT) approach to investigate the internal motivation of users topic preferences in online social media. Through the proposed three-level generation framework, the NHBT model identifies the multiple-to-multiple generative relationships from intrinsic factors to topic preferences. The NHBT model also allows us to identify the dominant intrinsic factors to explain individuals topic selection behaviors and envisage the intrinsic factors for new users. Our experiments show that the NHBT model is efficient to draw topics from online social media and reveal the intrinsic mixed mechanisms for users topic selection behaviors.

In terms of future research, literature shows that some words are usually significantly related to specific personality dimensions. A possible direction is to integrate useful prior information about the correlation between specific words and personality dimensions to improve the effect of NHBT model. In real context of online social media, topics usually co-occur with user sentiments. Another possible extension is to integrate user sentiments into NHBT model and design models to study the generative relationships from intrinsic factors to both topics and sentiments.

Acknowledgements

This work is supported by the Major Program of the National Natural Science Foundation of China (71490725), the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (71521001), the National Natural Science Foundation of China (71722010, 91546114, 71371062, 71302064, 71501057), the National Key Basic Research Program of China (2013CB329603), the National Key Technology Support Program (2015BAH26F00).

References

[1] M.H. Alam, W.J. Ryu, S.K. Lee, Joint multi-grain topic sentiment: modeling semantic aspects for online reviews, Inf. Sci. 339 (2016) 206–223.
[2] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[3] L. Cao, In-depth behavior understanding and use: the behavior informatics approach, Inf. Sci. 180 (17) (2010) 3067–3085.
[4] C.-C. Chang, C.-J. Lin, Libsvm: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.
[5] A. Chauhan, K. Kummamuru, D. Toshniwal, Prediction of places of visit using tweets, Knowl. Inf. Syst. 50 (1) (2017) 145–166.
[6] R. Cheng, B. Tang, A music recommendation system based on acoustic features and user personalities, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2016, pp. 203–213.
[7] T. Correa, A.W. Hinsley, Who Interacts on the Web?: The Intersection of Users' Personality and Social Media Use, Elsevier Science Publishers B. V., 2010.
[8] E. Erosheva, S. Fienberg, J. Lafferty, Mixed-membership models of scientific publications, Proc. Nat. Acad. Sci. 101 (suppl 1) (2004) 5220–5227.
[9] T.S. Ferguson, A bayesian analysis of some nonparametric problems, Ann. Stat. 1 (2) (1973) 209–230.
[10] K.R. Fricke, P.Y. Herzberg, Personality and self-reported preference for music genres and attributes in a german-speaking sample, J. Res. Pers. 68 (2017) 114–123.
[11] github, A java classifier based on the naive bayes approach complete with maven support and a runnable example (2012).
[12] J. Han, H. Lee, Characterizing the interests of social media users: refinement of a topic model for incorporating heterogeneous media, Inf. Sci. 358359 (2016) 112–128.
[13] K. Harris, T. English, P.D. Harms, J.J. Gross, J.J. Jackson, Why are extraverts more satisfied? personality, social experiences, and subjective wellbeing in college, Eur. J. Pers. 31 (2) (2017) 170–186.
[14] S. Jeon, S. Kim, H. Yu, Spoiler detection in TV program tweets, Elsevier Science Inc., 2016.
[15] Y. Jo, A.H. Oh, Aspect and sentiment unification model for online review analysis, in: ACM International Conference on Web Search and Data Mining, 2011, pp. 815–824.
[16] P. Kazienko, M. Adamski, Adrosaadaptive personalization of web advertising, Inf. Sci. 177 (11) (2007) 2269–2295.
[17] D. Kim, S. Kim, A. Oh, Dirichlet process with mixed random measures: a nonparametric topic model for labeled data, arXiv preprint arXiv:1206.4658 (2012).
[18] M. Kosinski, D. Stillwell, T. Graepel, Private traits and attributes are predictable from digital records of human behavior, Pnas 110 (15) (2013) 5802–5805.
[19] L. Liao, J. Jiang, Y. Ding, H. Huang, E.P. Lim, Lifetime lexical variation in social media, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014, pp. 1643–1649.
[20] J. Lin, W. Mao, D.D. Zeng, Personality-based refinement for sentiment classification in microblog, Knowl. Based Syst. 132 (2017) 204–214.
[21] J.D. Mcauliffe, D.M. Blei, Supervised topic models, in: Advances in neural information processing systems, 2008, pp. 121–128.
[22] M.R. Mehl, S.D. Gosling, J.W. Pennebaker, Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life, J. Personal. Social Psychol. 90 (5) (2006) 862.
[23] N.L. Muscanell, R.E. Guadagno, Make new friends or keep the old: gender and personality differences in social networking use, Comput. Human Behav. 28 (1) (2012) 107–112.
[24] R.M. Nallapati, A. Ahmed, E.P. Xing, W.W. Cohen, Joint latent topic models for text and citations, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, Usa, August, 2008, pp. 542–550.
[25] M.L. Newman, C.J. Groom, L.D. Handelman, J.W. Pennebaker, Gender differences in language use: an analysis of 14,000 text samples, Discourse Process. 45 (3) (2008) 211–236.
[26] G. Park, H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, M. Kosinski, D.J. Stillwell, L.H. Ungar, M.E.P. Seligman, Automatic personality assessment through social media language., J. Personal. Social Psychol. 108 (6) (2015) 934–952.
[27] J.W. Pennebaker, L.A. King, Linguistic styles: language use as an individual difference, J. Pers. Soc. Psychol. 77 (6) (1999) 1296–1312.
[28] M. Pita, M. Pita, A. Lacerda, A. Lacerda, G.L. Pappa, A general framework to expand short text for topic modeling, Inf. Sci. Int. J. 393 (C) (2017) 66–81.
[29] D. Ramage, D. Hall, R. Nallapati, C.D. Manning, Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora, in: Conference on Empirical Methods in Natural Language Processing: Volume, 2009, pp. 248–256.

[30] D. Ramage, C.D. Manning, S. Dumais, Partially labeled topic models for interpretable text mining, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, Ca, Usa, August, 2011, pp. 457–465.
[31] Y. Ren, Y. Wang, J. Zhu, Spectral learning for supervised topic models, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2016). 1–1.
[32] F. Rodrigues, M. Lourenco, B. Ribeiro, F. Pereira, Learning supervised topic models for classification and regression from crowds, IEEE Trans. Pattern Anal. Mach. Intell. PP (99) (2017). 1–1.
[33] C. Ross, E.S. Orr, M. Sisic, J.M. Arseneault, M.G. Simmering, R.R. Orr, Personality and motivations associated with facebook use, Comput. Human Behav. 25 (2) (2009) 578–586.
[34] H.A. Schwartz, J.C. Eichstaedt, M.L. Kern, L. Dziurzynski, S.M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M.E. Seligman, Personality, gender, and age in the language of social media: the open-vocabulary approach., Plos One 8 (9) (2013) e73791.
[35] J. Sethuraman, A constructive definition of dirichlet priors, Stat. Sin. (1994) 639–650.
[36] A. Srivastava, M. Sahami, Text Mining: Classification, Clustering, and Applications, 2009.
[37] Y.W. Teh, M.I. Jordan, M.J. Beal, D.M. Blei, Hierarchical dirichlet processes, J. Am. Stat. Assoc. 101 (476) (2006) 1566–1581.
[38] G.F. Templeton, X. Luo, T.R. Giberson, N. Campbell, Leader personal influences on membership decisions in moderated online social networking groups, Decis. Support Syst. 54 (1) (2012) 655–664.
[39] S. Tifferet, I. Vilnai-Yavetz, Gender differences in facebook self-presentation: an international randomized study, Comput. Human Behav. 35 (6) (2014) 388–399.
[40] H. Wang, F. Wu, W. Lu, Y. Yang, X. Li, X. Li, Y. Zhuang, Identifying objective and subjective words via topic modeling, IEEE Trans. Neural Netw. Learn. Syst. PP (99) (2017) 1–13.
[41] X. Wang, A. Mccallum, Topics over time: a non-markov continuous-time model of topical trends, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 424–433.
[42] Y.C. Wang, M. Burke, R.E. Kraut, Gender, topic, and audience response: An analysis of user-generated content on facebook, ACM, 2013.
[43] S. Wehrli, Personality on social network sites: an application of the five factor model, Eth Zurich Sociol. Working Pap. (2008).
[44] T. Yarkoni, Personality in 100,000 words: a large-scale analysis of personality and word use among bloggers, J. Res. Pers. 44 (3) (2010) 363–373.
[45] V. Zabkar, M. Arslanagic-Kalajdzic, A. Diamantopoulos, A. Florack, Brothers in blood, yet strangers to global brand purchase: a four-country study of the role of consumer personality (2017).
[46] Q. Zhang, L.T. Yang, Z. Chen, P. Li, High-order possibilistic c-means algorithms based on tensor decompositions for big data in iot, Inf. Fusion 39 (2018) 72–80.
[47] Q. Zhang, L.T. Yang, X. Liu, Z. Chen, P. Li, A tucker deep computation model for mobile multimedia feature learning, ACM Trans. Multimedia Comput., Commun., Appl. (TOMM) 13 (3s) (2017) 39.
[48] Q. Zhang, C. Zhu, L.T. Yang, Z. Chen, L. Zhao, P. Li, An incremental cfs algorithm for clustering large data in industrial internet of things, IEEE Trans. Ind. Inf. 13 (3) (2017) 1193–1201.