Learning Personalized Topical Compositions with Item Response Theory

Lu Lin
Department of Computer Science
University of Virginia
ll5fy@virginia.edu

Lin Gong
Department of Computer Science
University of Virginia
lg5bt@virginia.edu

Hongning Wang
Department of Computer Science
University of Virginia
hw5x@virginia.edu

ABSTRACT

A user-generated review document is a product between the item's intrinsic properties and the user's perceived composition of those properties. Without properly modeling and decoupling these two factors, one can hardly obtain any accurate user understanding nor item profiling from such user-generated data.

In this paper, we study a new text mining problem that aims at differentiating a user's subjective composition of topical content in his/her review document from the entity's intrinsic properties. Motivated by the *Item Response Theory (IRT)*, we model each review document as a user's detailed response to an item, and assume the response is jointly determined by the individuality of the user and the property of the item. We model the text-based response with a generative topic model, in which we characterize the items' properties and users' manifestations of them in a low-dimensional topic space. Via posterior inference, we separate and study these two components over a collection of review documents. Extensive experiments on two large collections of Amazon and Yelp review data verified the effectiveness of the proposed solution: it outperforms the state-of-art topic models with better predictive power in unseen documents, which is directly translated into improved performance in item recommendation and item summarization tasks.

CCS CONCEPTS

• **Information systems** → *Recommender systems*; Personalization; *Document topic models*; • **Mathematics of computing** → Probabilistic inference problems;

KEYWORDS

Correlated topic model; review mining; aspect modeling

ACM Reference Format:
Lu Lin, Lin Gong, and Hongning Wang. 2019. Learning Personalized Topical Compositions with Item Response Theory. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3289600.3291022

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3291022

1 INTRODUCTION

The explosive growth of social media empowers more and more people to freely express and share their opinions on all kinds of entities such as products and services. Among them, textual reviews, a typical type of user-generated content describing how a user evaluates different aspects of an entity, serve as an increasingly important proxy to understand users [13, 15, 27] and profile entities [19, 22, 28]. Although various research efforts have been devoted to this important area (see [20] for a comprehensive literature survey), one fundamental question remains largely unresolved: how to distinguish a user's subjective compositions of different aspects of an entity in his/her review comments from the entity's intrinsic properties? This question naturally emerges as various qualitative studies [21, 34] show that for the same item, different users might focus on different aspects of it in their review comments; and even for the same user towards a set of items of the same type, he/she might still cover different aspects in his/her reviews. This phenomenon cannot be simply explained by the heterogeneity in users' topical emphases, because some consistent patterns occur across users and items. As a result, user understanding and item profiling can hardly be accurate, until the mechanism underlying such variance is thoroughly explored and modeled.

The first step to study this problem is to quantify a user's topical emphasis in his/her generated review text content. Statistical topic models [3, 11] serve as a powerful tool to analyze such data, by which the learnt topical structure unveils users' attention over different aspects of entities. Topic models represent a document as a mixture of latent topics; and different types of generative assumptions have been imposed in learning the dependency among users, items and topics. Rosen-Zvi et al. [25] assume a multi-author document is a mixture of its authors' topical interests; by fitting the model over a corpus of documents, each author's topic distribution can be recovered. Xu et al. [32] consider different factors behind the generation of a user's social media posts, e.g., influence from friends v.s., personal interests, and use a topic model to estimate each of the factors. [13, 15, 27] study users' fine-grained sentiment evaluation of entities at the level of topical aspects. And with additional knowledge of user provided item ratings, [16, 26, 31] combine topic modeling with collaborative filtering to identify the latent representation for both users and items.

But none of the existing text modeling solutions differentiate a user's subjective composition of topical aspects in his/her review content from the entity's intrinsic properties. Nevertheless, these two factors are never separated when a user is creating his/her opinionated evaluations of a particular item. For example, when reviewing the same restaurant, one user might correlate its price aspect with service aspect, such that whenever he/she comments

on price, he/she will also comment on service; while another user might correlate price with location. As a result, when one reads the reviews from these two users, it is hard to tell if the restaurant is featured with service and location, or just has a competitive price. Like what is described in the famous parable of the blind men and an elephant, humans have a tendency to exploit and manipulate their partial experiences and bias when describing their observations.

To formally model the personalized topical compositions in review text data, we take a perspective stemmed from the *Item Response Theory* (IRT) [7, 24]. IRT was originally developed for psychometrics studies, and it states that the probability of a specific response is determined by the item's attribute and the subject's individuality. Both the item's attribute and the subject's individuality are considered as latent variables. Mapping it to the problem studied in this work, we consider the generation of a review document as a detailed user response towards a given item, where the user's topical composition (individuality) and the item's aspect-level property (attribute) jointly generate the review content (response). By inferring the posterior distribution of the latent variables over a set of review documents, we can distill the items' intrinsic aspect-level properties from the user's personalized topical compositions.

We realize this modeling principle with topic models. We assume items from the same category share the same set of topical aspects, which are modeled as a word distribution over a fixed vocabulary. Each item is characterized as a unique mixture over the topics, reflecting its aspect-level properties. Each user is modeled as a linear composition function, which transforms the item-level topic mixture to a document-level topic mixture. As a result, each review document is generated by sampling from the corresponding user-item specific topic mixture. We name the resulting model as Topical User-Item Response (TUIR) model. To decouple the interdependence between an item's intrinsic topical property and a user's personal topical composition, we apply a variational Bayesian method for efficient posterior inference. Extensive experimental evaluations on large-scale Amazon and Yelp review datasets demonstrate the unique value of the proposed solution: by decomposing the two factors, TUIR obtains better predictive power in unseen documents, especially those from known users or items, which is directly translated into improved quality in item recommendation and summarization.

In summary, our work makes the following contributions,

- We propose a new text mining problem of decoupling users' subjective topical compositions from items' intrinsic properties. It enables accurate user modeling and item profiling.
- Motivated by the Item Response Theory, we develop a novel probabilistic topic model, named Topical User-Item Response (TUIR) model, to address the decomposition problem. We provide a tight variational Bayesian solution to efficiently perform its posterior inference.
- We perform extensive empirical evaluations on real-world review datasets to investigate the value of the proposed decomposition problem and demonstrate its utility in applications of user modeling and item recommendation.

2 RELATED WORK

Earlier works in user-generated content modeling focus on word level analysis of text data, where user opinions [6, 30] and item features [12, 35] are summarized. Manually crafted lexicon features or frequent pattern mining methods are typically used to model the text content. Limited by the features' generalization capability, such methods can only provide population-level content analysis, e.g., summarizing all users' reviews about a particular item, and cannot reflect individual users' detailed evaluations.

Statistical topic models [3, 11] provide a more principled way in modeling text data. Classical topic models focus on document analysis, where each document is modeled as a mixture of topics and each topic is modeled as a distribution over words. Various topical structures have been proposed to model the correlation between topics [1], topic evolution over time [2], and hierarchical structure among topics [8]. Such generic purpose topic models are useful to discover hidden thematic structures in a corpus, but they do not explicitly model users or items. Ad-hoc post-processing is needed to obtain user-level and item-level analysis, e.g., aggregate document-level topics to users and items.

Many works extend classical topic models to capture topical dependency between users and documents. Author-Topic model [25] assumes the topic distribution in a multi-author document is a simple mixture of each author's topic distribution, so that the authors can be profiled in a topic space. Follow-up works extend it by introducing more detailed dependency assumption among the authors. For example, Author-Persona-Topic model [18] captures author's expertise by topical mixtures associated with each individual author; author-Recipient-Topic model [17] analyzes the direction-sensitive messages sent between individuals and learns topic distribution conditioned on both senders and recipients.

Topic models have also been developed to model users' opinionated assessment of items [13, 15, 27]. But most of such models focus on the dependency between topical aspects and sentiment, rather than that between users and items. For example, in [27], the researchers assume different users would weight different aspects differently when generating the overall opinion rating of an item, but the weights are assumed to be item independent. The most related works to ours are collaborative filtering based topic models [16, 26, 31], which map both users and items into a shared topic space for opinion rating prediction. However, the interaction between users and items are only modeled for interpreting the sentiment ratings; while the text content is still assumed to be solely dependent on the user or the item. In addition, the requirement on explicit opinion ratings also limits their applications in analyzing users and items when only rating data is available.

To briefly summarize, most of existing solutions do not consider the interdependence between users and items when modeling user-generated text content. They accredit the generation of documents to either users or corpus-wise prior, and therefore cannot well separate users' personal composition of topics from items' intrinsic property at the topic level.

3 TOPICAL USER-ITEM RESPONSE MODEL

In this section, we discuss our generative model motivated by the Item Response Theory, according to which we treat a review document as a user's detailed response to a specific item and assume the response is jointly determined by the *individuality* of the user and *attribute* of the item. We then present the technical details of our

model, including the procedures for model specification, posterior inference, and parameter estimation.

3.1 An Item Response Theory View of User-Generated Review Data

The Item Response Theory (IRT) [7, 24], also known as the latent trait model, refers to a family of psychometrics models that attempt to explain the relationship between latent traits (unobservable characteristic or attribute) and their manifestations (i.e., observed outcomes, responses or performance). They establish a link between the properties of items, individuals responding to these items, and the underlying trait being measured. IRT assumes that the latent construct of an individual (e.g. stress, knowledge, attitudes) and items of a measure are organized in an unobservable continuum. Therefore, its main purpose focuses on establishing the individual's position on that continuum. Although it was originally designed for psychometrics studies, e.g., design, analysis and scoring of tests and questionnaires, IRT has also been successfully applied in heath outcomes research [9] and personalized online education [5].

In this work, we treat a given review document as a user's response to an item. To realize IRT formulations in text data, we model each item's intrinsic property as a distribution over topics. The topic distribution thus uniquely characterizes the item's attribute at the level of aspects and attracts users' attention. When creating a response to an item, we assume a user would first examine the topic proportion of the item, map it to his/her personal composition of topics for this item, and then generate the review content accordingly. Each review document is thus modeled as a compound of these two factors; and via posterior inference over a set of documents, we estimate and decouple these two factors.

There are several key assumptions in IRT, which are naturally applied to our modeling of user-generated review content. First, it assumes the probability of a correct response increases with the trait level increasing. The traits being measured in IRT can be considered as the composed topics in a user review. When the proportion of a particular topic increases in a review document, it becomes more likely for us to observe words related to it in that document. Second, IRT assumes that there are only a finite number of dominant traits being measured (i.e., multi-dimensional IRT model [23]) and those traits are the driving force for the responses observed for each item in the measure. In our proposed solution, we model the latent attributes of items with topics and assume a fixed number of them in a collection of items. Third, responses given to the separate items in a test are mutually independent given a certain level of ability. This corresponds to that in our assumption given a pair of user and item, the generation of text content in one review is independent from other reviews. Fourth, one can estimate the parameters of an item from any group of individuals who have answered the item. This requires the item's attributes to be independent from the users who provide responses on it. And this is also the key assumption in our solution, where we consider an item's intrinsic properties and the way that a user composes the topics are independent from each other in prior.

Our solution also extends classical IRT. In all IRT models, only single dimensional responses are considered, such as Dichotomous IRT models for True/False questions and nominal response models for expected scores of testing items. In our work, we treat a review document as a user's response to an item, which is essentially a high dimensional word vector. We appeal to a generative modeling approach to formalize the generation of such vectors in a lower dimensional topic space, which summarizes the thematic patterns among words. Its immediate benefit is an interpretable representation of items' intrinsic proprieties and how different users perceive those proprieties when evaluating the items.

3.2 Model Specification

We first establish notations. In a collection of review documents about items from a particular category, we assume there is a set of users $\mathcal{U} = \{u_1, u_2, \ldots, u_U\}$ and a set of items $\mathcal{I} = \{i_1, i_2, \ldots, i_I\}$. Each user u provides reviews for a list of items $\mathcal{I}_u \subset \mathcal{I}$; and accordingly, each item i is reviewed by a list of users $\mathcal{U}_i \subset \mathcal{U}$. A review document $d_{i,u}$, which is considered as user u's response to item i, is denoted as a bag of words $d_{i,u} = (w_1, w_2, \ldots, w_N)$, where w_n is from a vocabulary of fixed size V. We assume K topics underlying the category of items, and each topic k is modeled as a multinomial distribution, i.e., Multi(β_k), over the vocabulary [3, 11]. Each item i is associated with a topic distribution quantified by the mixing proportion γ_i, which depicts the item's intrinsic properties in the topic space.

The key in modeling the generation of review documents is to specify the topic distribution in each individual document. Based on our view of user-generated review data from item response theory, we consider this document-level topic distribution as a mapped result from the item's intrinsic topical distribution to the user's personal composition of it. To realize this mapping, we introduce a topical composition matrix for each user u, $P_u \in \mathbb{R}^{K\times K}$, where each column k represents how user u correlates other topics with the kth topic in his/her view. Consequently, the inner product $P_{u,.,k}^\top \gamma_i$ measures user u's perception of item i on topic k, and it thus serves as a mapped topic distribution for the review document $d_{i,u}$.

We should note that to be qualified as a topic distribution, the vector $P_u^\top \gamma_i$ has to lie in a K-dimensional simplex. Instead of requiring each column of P_u to lie in the simplex (as γ_i is already in this simplex), we use a logistic-normal distribution to map $P_u^\top \gamma_i$ to the simplex [1]. The advantages are two folds: first, without restricting the sign of P_u, it is possible to learn negative correlation between topics, which increases the descriptive power of the model; second, it reduces the number of constraints to be imposed in posterior inference (we will elaborate on this later) and thus reduces the computational complexity. Denote the document-level topic vectors as $\theta_{i,u} \in \mathbb{R}^K$, we have $\theta_{i,u} \sim \mathcal{N}(P_u^\top \gamma_i, \rho^{-1}\boldsymbol{I})$; and according to $\theta_{i,u}$, we sample topic indicator $z_{i,u,n} \in \{1, \ldots, K\}$ for each word $w_{i,u,n}$ by $z_{i,u,n} \sim \text{Multi}(\text{softmax}(\theta_{i,u}))$ in the review document $d_{i,u}$.

To make our model a complete generative model, we also impose priors over the user-level topic composition matrix P_u and the item-level topic distribution γ_i. For each column k in P_u, we have $P_{u,.,k} \sim \mathcal{N}(\mathbf{0}, \sigma^{-1}\boldsymbol{I})$, and for each item i, we have $\gamma_i \sim \text{Dir}(\alpha)$, where σ and α are corresponding hyper-parameters. Putting these components together, the generative process of our solution can be described as follows,

- For each item i:
  - Draw item topic distribution $\gamma_i \sim \text{Dir}(\alpha)$
- For each user u:
  - For each topic $k = 1, 2, \ldots, K$:

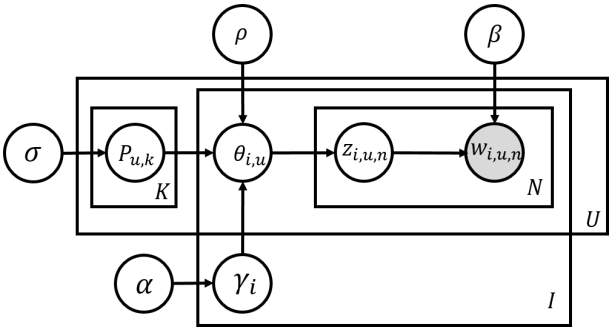

Figure 1: Graphical model presentation of the Topical User-Item Response model. In this model, each item's intrinsic aspect-level property is modeled as a topic mixture vector γ, and the user's topical composition is modeled as a linear transformation matrix P. The inner product between P and γ determines the topic proportion θ in each individual document from the user about the item.

    - Draw user topic composition vector $P_{u,.,k} \sim \mathcal{N}(\mathbf{0}, \sigma^{-1}\boldsymbol{I})$
- For each review document from user u about item i:
  - Draw user-item document topic vector $\theta_{i,u} \sim \mathcal{N}(P_u^\top \gamma_i, \rho^{-1}\boldsymbol{I})$
  - For each word $w_{i,u,n}$ in document $d_{i,u}$:
    - Draw topic assignment $z_{i,u,n} \sim \text{Multi}(\text{softmax}(\theta_{i,u}))$
    - Draw word $w_{i,u,n} \sim \text{Multi}(\beta_{z_{i,u,n}})$.

We name the resulting generative model as Topic User-Item Response Model, or TUIR in short. And the graphical model representation is provided in Figure 1.

As illustrated in the figure, a collection of user-generated reviews can be considered as a bipartite graph between users and items, with review documents as edges connecting them. On each edge, we have $\theta_{i,u} \sim \mathcal{N}(P_u^\top \gamma_i, \rho^{-1}\boldsymbol{I})$, which can be viewed as a multivariate multiple regression problem [14] from item i to user u's response. Specifically, γ_i is the predictor variable, $\theta_{i,u}$ is the response vector, and P_u is the regression coefficient. And we assume the error term in each regression problem follows a zero mean isometric Gaussian distribution. However, as the topic proportion vector $\theta_{i,u}$ in each document is not directly observable, we treat it as a latent variable and infer its posterior distribution based on the observed document content. As a result, each item's intrinsic topic distribution is learnt through the observed review content and the corresponding user's topic composition mapping; and respectively, each user's topic composition mapping is estimated based on his/her reviewed item's intrinsic topic distribution. Next, we will introduce an efficient variational Bayesian method to perform the posterior inference of the latent variables of interest; and based on it we apply the Expectation-Maximization algorithm to estimate the model parameters in TUIR.

3.3 Variational Bayesian for Posterior Inference

The latent variables of interest in TUIR are P, γ, θ and $\boldsymbol{z}$, which represent the user-level topic composition, item-level topic distribution, document-level topic proportion, and topic assignments of each word in a document. But because of the coupling among these latent variables, e.g., $\theta_{i,u}$ is determined by a linear composition between P_u and γ_i, exact posterior inference is intractable. We appeal to a variational Bayesian method for approximated inference.

The basic idea of variational inference is to exploit the convexity of the log-likelihood function to obtain and maximize a lower bound of it. Based on mean-field approximation, we drop the dependency among $P, \gamma, \theta, \boldsymbol{z}$ and $\boldsymbol{w}$ to obtain a family of simplified variational distributions for these latent variables. Each of these variational distributions is governed by their corresponding free vatiational parameters:

$$q(P, \gamma, \theta, \boldsymbol{z}|\eta, \nu, \Sigma^{(p)}, \mu, \Sigma^{(\theta)}, \phi) = \prod_{i=1}^{I} q(\gamma_i|\eta_i) \prod_{u=1}^{U}\prod_{k=1}^{K} q(P_{u,.,k}|\nu_{u,k}, \Sigma^{(p)}_{u,k}) \\ q(\theta_{i,u}|\mu_{i,u}, \Sigma^{(\theta)}_{i,u}) \prod_{n=1}^{N} q(z_{i,u,n}|\phi_{i,u,n})$$

where $q(\gamma_i|\eta_i)$ follows a Dirichlet distribution parameterized by η_i, $q(P_{u,.,k}|\nu_{u,k}, \Sigma^{(p)}_{u,k})$ follows a multivariate Gaussian distribution with mean $\nu_{u,k}$ and covariance $\Sigma^{(p)}_{u,k}$, $q(\theta_{i,u}|\mu_{i,u}, \Sigma^{(\theta)}_{i,u})$ follows a multivariate Gaussian distribution with mean $\mu_{i,u}$ and variance $\Sigma^{(\theta)}_{i,u}$, and $q(z_{i,u,n}|\phi_{i,u,n})$ follows a Multinomial distribution parameterized by $\phi_{i,u,n}$. Because the topic proportion vector $\theta_{i,u}$ is inferred in each document, it is not meaningful to estimate a full covariance matrix for it [1]. Hence, in its variational distribution, we only estimate the variance parameters.

Based on the introduced variational distributions, the log-likelihood of a review document is then lower bounded by Jensen's inequality:

$$\begin{aligned} &\log p(\boldsymbol{w}|\alpha, \beta, \sigma, \rho) \\ =&\log \int\int\int \sum_{z} \frac{p(P, \gamma, \theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta, \sigma, \rho) q(P, \gamma, \theta, \boldsymbol{z})}{q(P, \gamma, \theta, \boldsymbol{z})} dP d\gamma d\theta \\ \geq&\mathbb{E}_q\big[\log p(P, \gamma, \theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta, \sigma, \rho)\big] - \mathbb{E}_q\big[\log q(P, \gamma, \theta, \boldsymbol{z})\big] \end{aligned} \quad (1)$$

Let $\mathcal{L}(\eta, \nu, \Sigma^{(p)}, \mu, \Sigma^{(\theta)}, \phi; \alpha, \beta, \sigma, \rho)$ denote the right-hand side of Eq (1). We can easily verify [3],

$$\begin{aligned} &\mathbb{D}_{KL}\big(q(P, \gamma, \theta, \boldsymbol{z}) \| p(P, \gamma, \theta, \boldsymbol{z})\big) \\ =&\mathbb{E}_q[\log q(P, \gamma, \theta, \boldsymbol{z})] - \mathbb{E}_q[\log p(P, \gamma, \theta, \boldsymbol{z})] \\ =&\mathbb{E}_q[\log q(P, \gamma, \theta, \boldsymbol{z})] - \mathbb{E}_q[\log p(P, \gamma, \theta, \boldsymbol{z}, \boldsymbol{w}) - \log p(\boldsymbol{w}|\alpha, \beta, \sigma, \rho)] \\ =&\mathbb{E}_q[\log p(\boldsymbol{w}|\alpha, \beta, \sigma, \rho)] - (\mathbb{E}_q[\log p(P, \gamma, \theta, \boldsymbol{z}, \boldsymbol{w})] - \mathbb{E}_q[\log q(P, \gamma, \theta, \boldsymbol{z})]) \\ =&\log p(\boldsymbol{w}|\alpha, \beta, \sigma, \rho) - \mathcal{L}(\eta, \nu, \Sigma^{(p)}, \mu, \Sigma^{(\theta)}, \phi; \alpha, \beta, \sigma, \rho) \end{aligned}$$

which suggests that minimizing the KL divergence between the variational posterior and the true posterior is equivalent to maximizing the lower bound of data likelihood with respect to the free variational parameters $(\eta, \nu, \Sigma^{(p)}, \mu, \Sigma^{(\theta)}, \phi)$.

The first step to maximize this lower bound is to derive the analytic form of posterior expectations required in $\mathcal{L}(\eta, \nu, \Sigma^{(p)}, \mu, \Sigma^{(\theta)}, \phi; \alpha, \beta, \sigma, \rho)$. As we have introduced conjugate priors for $\{P_u\}_{u=1}^U$ and $\{\gamma_i\}_{i=1}^I$, the expectations related to these latent variables have closed form solutions. But as there is no conjugate prior for logistic Normal distribution, we apply variational inference to approximate the expectations related to $\theta_{u,i}$. Next we describe the detailed inference procedure for each latent variable, and due to space limit we will omit most of details for the expectation calculation.

• **Estimate item topic distribution parameter η.** For each item i, we relate the terms associated with $q(\gamma_i|\eta_i)$ in Eq (1) to form a

function $\mathcal{L}_{[\eta_i]}$, and maximize it to estimate η_i,

$$\mathcal{L}_{[\eta_i]} = \sum_{k=1}^{K}(\alpha_k - \eta_{i,k})\big(\Psi(\eta_{i,k}) - \Psi(\sum_{j=1}^{K}\eta_{i,j})\big) \tag{2}$$
$$+ \sum_{k=1}^{K}\log\Gamma(\eta_{i,k}) - \log\Gamma(\sum_{k=1}^{K}\eta_{i,k}) + \frac{\rho}{\eta_{i,0}}\sum_{u=1}^{U_i}\sum_{j=1}^{K}\sum_{k=1}^{K}\eta_{i,k}\nu_{u,j,k}\mu_{i,u,j}$$
$$- \frac{\rho}{2\eta_{i,0}(\eta_{i,0}+1)}\sum_{u=1}^{U_i}\sum_{j=1}^{K}\sum_{k=1}^{K}\Big(\sum_{l=1}^{K}\eta_{i,l}\eta_{i,k}(\Sigma^{(p)}_{u,j,l,k} + \nu_{u,j,l}\nu_{u,j,k})$$
$$+ \sum_{l=k}^{K}\eta_{i,l}(\Sigma^{(p)}_{u,j,l,k} + \nu_{u,j,l}\nu_{u,j,k})\Big)$$

where $\Psi(\cdot)$ is the first order derivative of log Gamma function. Since there is no close-form solution of this optimization problem, we use gradient ascent to iteratively estimate it. Due to space limit, we will omit the derived gradient; but the meaning behind this optimization problem for η_i is self-evident. The first part of Eq (2) represents the regularization of η_i from its prior distribution $\mathrm{Dir}(\alpha)$ (e.g., the summation over K topics), and the second part of it represents the loss from regression programs from γ_i to $\theta_{u,i}$ over all reviews talking about item i (e.g., the summation over U_i).

• **Estimate user topic composition parameter $\nu, \Sigma^{(p)}$.** For each user u, we relate the terms that are associated with $q(P_u|\nu_u, \Sigma^{(p)}_u)$ to form the objective function $\mathcal{L}_{[\nu,\Sigma^{(p)}]}$, and estimate the variational parameters by maximizing this function. Fortunately, closed form estimation of $\nu, \Sigma^{(p)}$ exists,

$$\nu_{u,k} = \rho\Sigma^{(p)}_{u,k}\sum_{i=1}^{I_u}\frac{\eta_i}{\eta_{i,0}}\mu_{i,u,k} \tag{3}$$

$$\Sigma^{(p)}_{u,k} = \Big(\sigma\boldsymbol{I} + \rho\sum_{i=1}^{I_u}\frac{\eta_i\eta_i^\top + \mathrm{diag}(\eta_i)}{\eta_{i,0}(\eta_{i,0}+1)}\Big)^{-1} \tag{4}$$

Note that the estimation of $\Sigma^{(p)}_{u,k}$ is not related to k, because we impose an isometric Gaussian prior for P_u in TUIR. This suggests that we implicitly assume the correlations between the composition coefficients across the topics in each user are homogeneous. The meaning behind this estimation is also clear: As we can consider user u's topic composition matrix P_u as a regression coefficient between item-level and document-level distributions, and we have assumed a zero mean Gaussian distribution for the regression error term, estimations of ν and $\Sigma^{(p)}$ are basically the mean and covariance estimations of P_u under the variational distribution.

• **Estimate document topic proportion parameter $\mu, \Sigma^{(\theta)}$.** Similar procedures as above can be taken to estimate these two variational parameters. However, since a logistic normal distribution does not have a conjugate prior, we again apply variational inference for it by introducing an additional free variational parameter ζ in each document. Because there is no closed form solution for the resulting optimization problem, we use gradient ascent to optimize μ and $\Sigma^{(\theta)}$ with the following gradients,

$$\frac{\partial\mathcal{L}}{\partial\mu_{i,u,k}} = -\rho(\mu_{i,u,k} - \frac{\eta_i^\top}{\eta_{i,0}}\nu_{u,k}) + \sum_{n=1}^{N}\phi_{i,u,n,k} \tag{5}$$
$$- N\zeta_{i,u}^{-1}\exp(\mu_{i,u,k} + \frac{1}{2}\Sigma^{(\theta)}_{i,u,k,k})$$

$$\frac{\partial\mathcal{L}}{\partial\Sigma^{(\theta)}_{i,u,k,k}} = \frac{1}{2}\Big(\frac{1}{\Sigma^{(\theta)}_{i,u,k,k}} - \rho - N\zeta_{i,u}^{-1}\exp(\mu_{i,u,k} + \frac{1}{2}\Sigma^{(\theta)}_{i,u,k,k})\Big) \tag{6}$$

where $\zeta_{i,u} = \sum_{k=1}^{K}\exp(\mu_{i,u,k} + \frac{1}{2}\Sigma^{(\theta)}_{i,u,k,k})$. As we mentioned before, because only the diagonal elements in $\Sigma^{(\theta)}_{i,u}$ are statistically meaningful (i.e., variance), we simply set its off-diagonal elements to zero. As the variance has to be non-negative, we can instead estimate the square root of it to avoid solving a constrained optimization problem. The gradient function suggests that the document-level topic proportion vector should align with the projected item topic distribution by this user and the inferred topic distribution based on its document content. This corresponds to our multivariate multiple regression interpretation of our developed model.

• **Estimate word topic assignment parameter ϕ.** This variational parameter can be easily estimated by $\phi_{i,u,n,k} \propto \exp(\mu_{i,u,k} + \sum_{v=1}^{V}w_{i,u,n,v}\log\beta_{k,v})$ for each individual word in each review document.

The above variational inference procedures are executed in an alternative fashion until the lower bound converges. Because the variational parameters can be grouped into user-level (ν and $\Sigma^{(p)}$), item-level (η), and document-level ($\mu, \Sigma^{(\theta)}$ and ϕ) parameters, the alternative update can be performed in parallel to improve efficiency. For example, fix $\nu, \Sigma^{(p)}$ and η, and distribute the documents across different machines to estimate their own $\mu, \Sigma^{(\theta)}$ and ϕ in parallel for large collections of documents.

3.4 Parameter Estimation

When performing the variational inference described above, we assume the model parameters α, β, σ and ρ are known ahead of time. But in practice, we also need to estimate them for a new collection of review documents. Based on the inferred posterior distribution of latent variables in TUIR, the model parameters can be readily estimated by the Expectation-Maximization (EM) algorithm.

Among the four model parameters, the most important ones are the prior for item-level topic distribution α and word-topic distribution β. As σ and ρ are two scalars serving as the variance for user topic composition matrix P_u and document topic proportion vector $\theta_{i,u}$, the model is less sensitive to their settings. Therefore, we will estimate α and β with respect to available training data, and empirically tune σ and ρ.

As there is no closed form solution for α with respect to Eq (1), we use gradient ascend for it with the following gradients,

$$\frac{\partial\mathcal{L}}{\partial\alpha_k} = I\big(\psi(\sum_{j=1}^{K}\alpha_j) - \psi(\alpha_k)\big) + \sum_{i=1}^{I}\big(\psi(\eta_{i,k}) - \psi(\sum_{j=1}^{K}\eta_{i,j})\big) \tag{7}$$

And the closed form estimation of β can be easily derived as,

$$\beta_{k,v} \propto \sum_{i=1}^{I}\sum_{u=1}^{U_i}\sum_{n=1}^{N}\phi_{i,u,n,k}w_{i,u,n,v} \tag{8}$$

where $w_{i,u,n,v}$ indicates the nth word in document $d_{i,u}$ is the vth word in the vocabulary.

In the resulting EM algorithm, in E-step variational inference procedures developed in Section 3.3 are executed until convergence; and in M-step α and β are estimated based on collected sufficient statistics from E-step. These two steps are iterated until the lower bound function $\mathcal{L}(\eta, \nu, \Sigma^{(p)}, \mu, \Sigma^{(\theta)}, \phi; \alpha, \beta, \sigma, \rho)$ converges over all training documents.

4 EXPERIMENTS

In this section, we evaluated the proposed TUIR model on two large collections of Amazon and Yelp review data. Extensive quantitative comparisons against several state-of-the-art models confirm the effectiveness of the proposed method. The tasks enabled by TUIR, including document modeling, collaborative filtering based item recommendation, and item summarization, further verified the effectiveness of the proposed model. Qualitative evaluation is performed to indicate the model's quality of representing users' personalized topical compositions and items' intrinsic properties based on the opinionated review text content.

4.1 Experimental Setup

Datasets. The evaluations are performed on two textual review corpora: 1) Restaurants, collected from Yelp [33], is a collection of Yelp restaurant reviews by removing users and restaurants with less than 15 reviews, which results in 78,462 reviews from 1,409 users and 843 restaurants; 2) Movies, collected from Amazon [10], is a collection of Amazon movie reviews by filtering users and movies with less than 40 reviews, which results in 71,174 reviews from 841 users and 919 movies. The filtering process is to obtain a dense user-item bipartite, and the threshold is determined by the average size of reviews per user. We selected 6,134 and 7,680 unigram and bigram textual features using the Information Gain (IG) selection method from the two datasets respectively. For both datasets, we randomly split the data for 5-fold cross validation in all the reported experiments.

Baselines. The proposed TUIR is compared against a rich set of baseline models, including classic topic modeling solutions and their user-specific or item-specific variants. 1) **Latent Dirichlet Allocation (LDA)** [3], it uses a conjugate Dirichlet prior distribution to model the topic distribution in documents. We extended two variants of it to fit the scenario of discovering user and item topical representations, by assuming different Dirichlet priors over different users or items. The variant considering user-specific Dirichlet prior is denoted as **uLDA**, and the other with item-specific Dirichlet prior is named as **iLDA**. 2) **Correlated Topic Model (CTM)** [1] employs a standard logistic-normal prior to capture pairwise topic correlations. 3) **Relational Topic Model (RTM)** [4] models the connection between two documents with a binary random variable conditioned on their textual contents. In our scenario, the document connection is defined by whether these two review documents belong to the same user or being associated with the same item. Therefore, RTM can be naturally extended to two variants, the one modeling documents linked by the same user is denoted as **uRTM**, and the other modeling documents linked by the same item is denoted as **iRTM**. We also included two sub-models of TUIR as baselines to evaluate the necessity of its user and item modeling components. By disabling the update of user topic compositions and setting it to an identity matrix, TUIR can only model items, thus is denoted as **iTUIR**. Correspondingly, TUIR with disabled update of item topic distribution is denoted as **uTUIR**. Note that among the baselines, LDA and CTM consider neither user or item factors; uLDA, uRTM and uTUIR only consider user factor; iLDA, iRTM and iTUIR only consider item factor; only TUIR inherently considers both user and item.

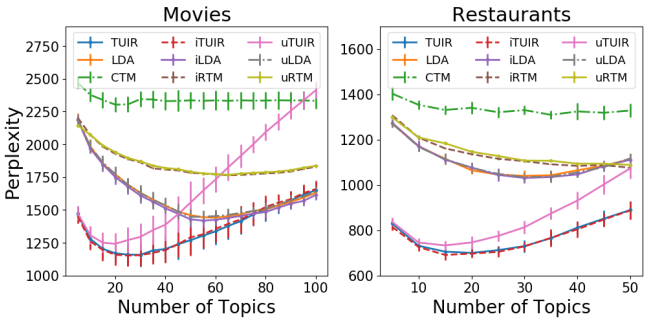

Figure 2: Comparison in perplexity on Movies and Restaurants.

4.2 Document Modeling

We first compare the predictive power of TUIR against baselines in the document modeling task. We compared all the topic models by their *perplexity* on the held-out test set to evaluate how likely the model will generate the observed text content in the held-out set. Formally, the perplexity for a set of unseen documents is calculated as follows [3]:

$$perplexity(D_{test}) = \exp\left(-\frac{\sum_{d \in D_{test}} \log p(\boldsymbol{w}_d)}{\sum_{d \in D_{test}} |d|}\right)$$

where the probability $p(\boldsymbol{w}_d)$ is estimated from the test set given a trained model. A better model on the document corpora will assign a higher probability to the held-out test set, and thus gives a lower perplexity score.

Figure 2 shows the mean and variance of perplexity for each model from 5-fold cross validation. TUIR achieves consistently the best predictive power on the hold-out dataset against all other topic models. This clearly demonstrates the advantage of the unique dependency structure imposed in TUIR, i.e., differentiating items' intrinsic topical properties from users' personal topic compositions, for modeling the review document content. Next, we will zoom into different sub-groups of hold-out datasets (i.e., reviews from seen users/items v.s., unseen users/items) to better illustrate the value of this imposed model structure in predicting unseen document content. In addition, Figure 2 also shows that TUIR's predictive capability peaks with a medium number of topics. This is expected: a smaller number of topics limits the model's resolution in recognizing the topical content, while a larger number of topics quickly increases the model's parameter size that calls for more training data. Both cases result in poorer performance of TUIR in modeling the user-generated text data.

To better understand the model's advantages in modeling the review data, we further segmented the held-out dataset into four subsets, considering whether the associated user or the associated item of a document in the held-out set appears in the training set,

- $D_{u\&i}^{Cold}$: Documents in which both the associated user and item do not occur in the training set, and thus the parameters for them are never learnt;
- D_{u}^{Cold}: Documents in which the associated item occur but the associated user does not occurred in the training set;
- D_{i}^{Cold}: Documents in which the associated user occur but the associated item does not occurred in the training set;

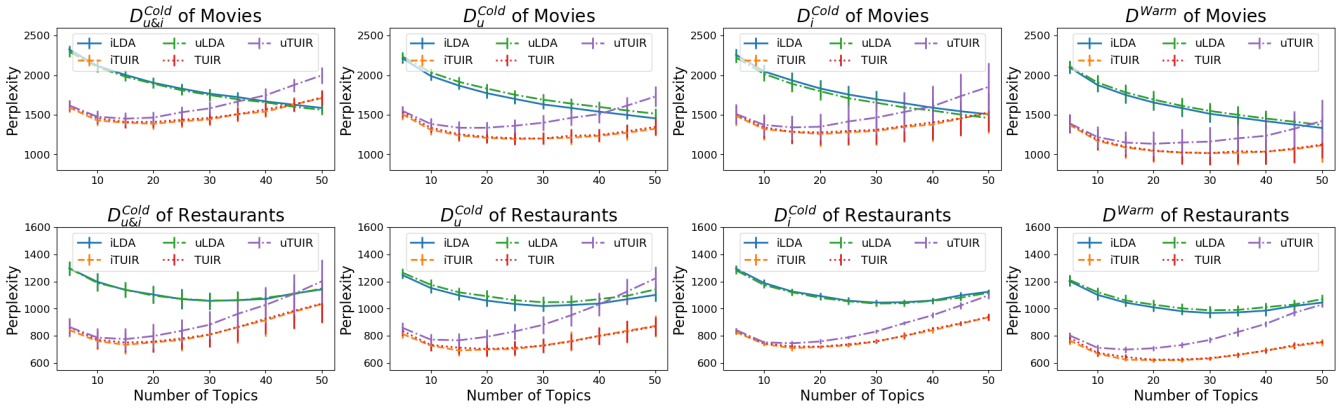

Figure 3: Perplexity results on Movies (Top) and Restaurants (Bottom) corpora in different subset of testing documents.

- D^{Warm}: Documents in which both the associated user and item occur in the training set, and thus the parameters for them are already learnt.

Documents in $D^{Cold}_{u\&i}$, D^{Cold}_{u} and D^{Cold}_{i} would suffer from cold start issue to different extends; which makes the accurate modeling of users and items essential in these settings. The calculation of perplexity is then performed in these four subsets respectively, and the results are reported in Figure 3. The greatly improved perplexity of TUIR on D^{Warm} compared with $D^{Cold}_{u\&i}$ indicates that the learnt item-level topic distribution and user-level topic composition are accurate and helpful for predicting future content from the known users about these observed items. In the meanwhile, by comparing the perplexity on D^{Cold}_{i} and D^{Cold}_{u}, we can observe that iTUIR generally performs better than uTUIR. This suggests the necessity of modeling individual items, since an item's intrinsic property is more distinctive for statistical learning and thus can largely help predict the future review content about it. Combing user's topic composition with iTUIR, a.k.a., TUIR, will further improve its generalization performance. Suffered from the cold start issue, the baseline models achieved much worse performance, which implies that the learnt TUIR model from known users and items serve as effective priors to predict text content in unseen users and items.

In order to illustrate the generative process modeled in TUIR where a review document is treated as a user's detailed response towards a given item, we estimate the TUIR model with 20 topics on a subset of Movies with 242 users and 134 movies, where the connection between users and items is denser. Figure 4 shows an example of how users respond to items on this dataset. In the meanwhile, we list the top words of representative topics learnt on this dataset and also on Restaurants dataset in Table 1.

In Figure 4, we illustrate three users along with their partial topic composition matrix P, and three movies with topic distribution vector γ. The edge denotes a review document, represented by the learnt topic proportion vector θ. The topic indices in this graph correspond to those in Table 1 (left). Darker color indicates a larger value. We can find in the result that u_1 has stronger emphasis in topic seven, thus the reviews of u_1 shift to the his/her topic interest. Movie i_3 is mostly featured by topic three, and thus user u_2 and u_3 comment on it with a larger coverage on this particular topic.

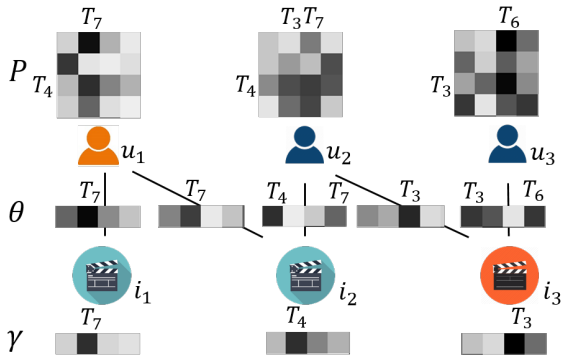

Figure 4: A portion of user-item bipartite graph where user's personalized topic composition matrix P and item's topic mixture γ are learned from Movies corpora. The edge between user and item represents corresponding review document, which is denoted by the topic proportion θ. Topic is denoted by T with index correspond to Table 1 (left).

This example reflects our intuition that user review content is a response of user's topic composition towards item's topic mixture.

4.3 Applications of TUIR

This section focuses on demonstrating the utility of users' topic compositions and items' topic-level aspects learnt by TUIR on two important application scenarios, collaborative filtering and item summarization.

• **Collaborative Filtering.** An essential question in Collaborative Filtering (CF) is how to estimate the similarity between users and items. Since TUIR simultaneously models users as topic compositions and items as topic mixtures, CF becomes an immediate application for TUIR to represent users and items at a topic level.

The collaborative filtering task is performed on the Movies dataset, where users review movies and provide star ratings from 1 to 5. We will first predict rating scores of a set of candidate movies given by a user, and the recommendation to this user is made by the ranking of predicted ratings. In our cross-validation, we obtained topical representations of users and items only on the training data

Table 1: Top words of example topics learnt by TUIR on Movies (left) and Restaurants (right) with 20 topics.

Topic	Top words
1	bond, bournj, jame, daniel, royal, queen, casino
2	futur, termin, predat, cameron, robot, magician, avatar
3	anim, wall, robot, pixar, nemo, superhero
4	xmen, battl, hels, cure, monster, spartan
5	thriller, villag, daughter, shyamalan, night, suspens
6	pirat, depp, jack, johnni, sparrow, bloom, captain
7	bruc, wayn, knight, citi, gotham, bale

Topic	Top words
1	pork, ramen, rice, thai, chicken, dish, beef
2	pizza, sauc, pasta, italian, crust, slice, garlic, bread
3	cream, chocol, cake, ice-cream, flavor, sweet, cooki,pastri
4	tabl, time, manag, NUM-minut, wait, restaur, seat
5	club, drink, night, peopl, music, vega, crowd, danc
6	sushi, roll, tuna, fish, salmon, rice, sashimi, sauc
7	taco, mexican, chip, bean, burrito, tortilla

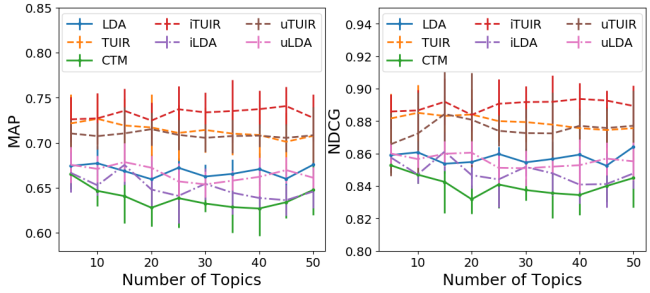

Figure 5: Collaborative filtering results of MAP (left) and NDCG (right) on Movies.

set. In TUIR, iTUIR and uTUIR, a user is modeled as the learnt topic composition matrices P_u, and an item is modeled as topic mixture vectors γ_i. In uLDA, users are represented by user-specific Dirichlet parameters. As in LDA, CTM and iLDA, no variables are specifically designed for users, we aggregate reviews from a user and average the posterior of document topic proportions as his/her user profile, which is denoted as $\bar{\theta}_u$ for user u. In testing, the rating of item i given by user u is predicted by the weighted average of observed ratings given by other users, and TUIR calculates the weight by the cosine similarity between document-level topic proportions, formally defined as follows:

$$w(\hat{d}_{i,u'}, d^*_{i,u}) = cosine(\hat{\theta}_{d_{i,u'}}, \text{softmax}(P_u^\top \gamma_i))$$

where $\hat{d}_{i,u'}$ is the review document of this item given by other users u' in the training set, and this document can be represented by the posterior topic proportion $\hat{\theta}_{i,u'}$ learnt from topic models. $d^*_{i,u}$ is the review document of this item given by a testing user u, and this review is unobserved along with the rating to be predicted; we use the inner product of the user's topic composition and the item's topic proportion with softmax as a proxy to predict the topic proportion in $d^*_{i,u}$. We call this setting user-based content collaborative filtering, since it estimates the content similarity across different users. As for other baselines, the weight is calculated by:

$$w(\hat{d}_{i,u'}, d^*_{i,u}) = cosine(\hat{\theta}_{d_{i,u'}}, \bar{\theta}_u)$$

where the topic proportion of unseen reviews is estimated by the user's topic mixture $\bar{\theta}_u$, because this is the only information baselines can provide to predict the content of $d^*_{i,u}$.

Given the weights, the candidate items to recommend is ranked by the predicted rating, which is the weighted average of ratings from observed user, and the ranking serves as our recommendation priority to the user. We evaluate the recommendation performance with Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP), and report the results in Figure 5. TUIR, together with iTUIR and uTUIR, outperformed others by 5% in MAP and 2.5% in NDCG, which suggests that user and item should be considered simultaneously, and purely using similarity between users is not sufficient to get the whole picture. This result also corresponds with the item response theory: rating acting as a response of user's personal interest towards item's intrinsic properties, it is not unilaterally determined by user or item.

• **Item Summarization.** We evaluate TUIR on item summarization task using the Restaurants dataset, where each restaurant has several official tags (short phrases) provided by Yelp to summarize the attributes of it. Since each item is usually reviewed by many users, one natural question is whether one can automatically summarize reviews regarding this item and generate "word-of-mouth" tags. This problem can be formulated as a ranking problem in information retrieval (IR), where the union of official tags across items is the ranking candidate pool, and the review set of a particular item serves as a query. Therefore, the item summarization problem can be rephrased as given an item with its reviews as query, finding top-k most likely tags from tag set for this item.

Two ranking methods are introduced as baselines: 1) Unigram **Language Model (LM)**, which calculates the likelihood of item reviews as the ranking score, given word probability learnt from each tag; 2) Okapi **BM25**, which calculates the score based on Term Frequency (TF) and Inverse Document Frequency (IDF). Correspondingly, the likelihood of each tag given an item can be directly calculated by topic models: $p(tag|item) = \prod_{w \in tag} \sum_z p(w|z)p(z|item)$, where $p(w|z)$ is the learnt topic-word distribution β, and $p(z|item)$ is the item's posterior topic mixture γ in TUIR, or aggregated average of topic mixture for reviews of this item $\hat{\theta}_i$ in CTM and LDA. We also adopt the model interpolation method introduced in [29] to combine BM25 with topic models, and name them with suffix "+BM25".

We rank tags with the ranking scores calculated by different models, and compare the rank with official tag set. The results of MAP and Mean Reciprocal Rank (MRR) are reported in Figure 6. BM25 outperforms others in MAP, while TUIR gives the best first-hit position evaluated by MRR. When combine BM25 with TUIR, the best MAP is achieved while maintaining a competitive MRR with TUIR. Large variance in CTM and LDA indicates the instability of representing items using aggregated document-level topic since it is heavily depended on single review document, which may inject too much user's personal bias. This result demonstrates the ability of TUIR in modeling items with a compact topic mixture while differentiating user's personal topical bias.

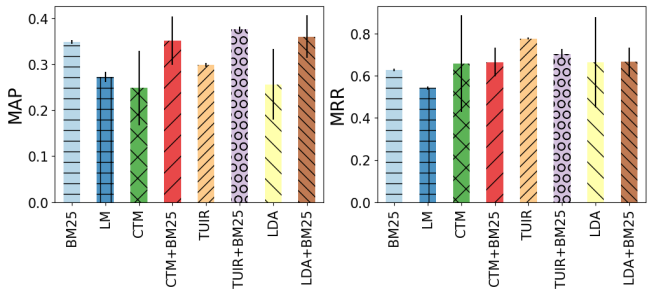

Figure 6: Item summarization results of MAP (left) and MRR (right) on Restaurants.

5 CONCLUSIONS AND FUTURE WORK

We studied a new text mining problem in this paper, which aims at differentiating a user's subjective composition of topical content in his/her review document from the entity's intrinsic properties. We developed a novel probabilistic topic modeling method motivated by the item response theory to address the problem. Extensive evaluations in document modeling, collaborative filtering and item summarization demonstrate the predictive power of our model, especially in new documents provided by existing users or items.

In our current model, we separately modeled users and items, ignoring the relatedness within users and among items. For example, similar users might share the same patterns in composing topics, and related items might share similar topical proprieties. It would be interesting to incorporate such relational information into our model. In addition, our model can be further extended to incorporate rating information for aspect-based sentiment analysis task.

6 ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments. This paper is based upon work supported by the National Science Foundation under grant IIS-1553568, IIS-1618948 and IIS-1718216.

REFERENCES

[1] David M Blei and John D Lafferty. 2005. Correlated topic models. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*. MIT Press, 147–154.
[2] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 113–120.
[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[4] Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *Artificial Intelligence and Statistics*. 81–88.
[5] Chih-Ming Chen, Hahn-Ming Lee, and Ya-Hui Chen. 2005. Personalized e-learning system using item response theory. *Computers & Education* 44, 3 (2005), 237–255.
[6] Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*. ACM, 519–528.
[7] Susan E Embretson and Steven P Reise. 2013. *Item response theory*. Psychology Press.
[8] Thomas L Griffiths, Michael I Jordan, Joshua B Tenenbaum, and David M Blei. 2004. Hierarchical topic models and the nested chinese restaurant process. In *Advances in neural information processing systems*. 17–24.
[9] Ron D Hays, Leo S Morales, and Steve P Reise. 2000. Item response theory and health outcomes measurement in the 21st century. *Medical care* 38, 9 Suppl (2000), II28.
[10] Ruining He and Julian McAuley. 2016. Ups and Downs: Modeling the Visual Evolution of Fashion Trends with One-Class Collaborative Filtering. *CoRR* abs/1602.01585 (2016). arXiv:1602.01585 http://arxiv.org/abs/1602.01585
[11] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 289–296.
[12] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 168–177.
[13] Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 815–824.
[14] Richard A Johnson and Dean W Wichern. 2004. Multivariate analysis. *Encyclopedia of Statistical Sciences* 8 (2004).
[15] Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 375–384.
[16] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. ACM, 165–172.
[17] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. 2005. Topic and Role Discovery in Social Networks.. In *Ijcai*, Vol. 5. Citeseer, 786–791.
[18] David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 500–509.
[19] Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 271.
[20] Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1–2 (2008), 1–135.
[21] Viktor Pekar and Shiyan Ou. 2008. Discovery of subjective evaluations of product features in hotel reviews. *Journal of Vacation Marketing* 14, 2 (2008), 145–155.
[22] Ana-Maria Popescu and Orena Etzioni. 2007. Extracting product features and opinions from reviews. In *Natural language processing and text mining*. Springer, 9–28.
[23] Mark D Reckase. 2009. Multidimensional item response theory models. In *Multidimensional Item Response Theory*. Springer, 79–112.
[24] Christine A Reid, Stephanie A Kolakowsky-Hayner, Allen N Lewis, and Amy J Armstrong. 2007. Modern psychometric methodology: Applications of item response theory. *Rehabilitation Counseling Bulletin* 50, 3 (2007), 177–188.
[25] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The Author-topic Model for Authors and Documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI '04)*. AUAI Press, Arlington, Virginia, United States, 487–494. http://dl.acm.org/citation.cfm?id=1036843.1036902
[26] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 448–456.
[27] Hongning Wang, Yue Lu, and ChengXiang Zhai. 2011. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 618–626.
[28] Nan Wang, Hongning Wang, Yiling Jia, and Yue Yin. 2018. Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. In *The 41st International ACM SIGIR Conference on Research, Development in Information Retrieval*. ACM, 165–174.
[29] Xing Wei and W. Bruce Croft. 2006. LDA-based Document Models for Ad-hoc Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 178–185. https://doi.org/10.1145/1148170.1148204
[30] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics, 347–354.
[31] Yao Wu and Martin Ester. 2015. Flame: A probabilistic model combining aspect based opinion mining and collaborative filtering. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 199–208.
[32] Zhiheng Xu, Yang Zhang, Yao Wu, and Qing Yang. 2012. Modeling user posting behavior on social media. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 545–554.
[33] Yelp. 2018. Yelp Dataset Challenge. https://www.yelp.com/dataset/challenge
[34] Xianfeng Zhang, Yang Yu, Hongxiu Li, and Zhangxi Lin. 2016. Sentimental interplay between structured and unstructured user-generated contents: an empirical study on online hotel reviews. *Online Information Review* 40, 1 (2016), 119–145.
[35] Li Zhuang, Feng Jing, and Xiao-Yan Zhu. 2006. Movie review mining and summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, 43–50.