Coupled Topic Model for Collaborative Filtering With User-Generated Content

Shu Wu, *Member, IEEE*, Weiyu Guo, Song Xu, Yongzhen Huang, *Member, IEEE*,
Liang Wang, *Senior Member, IEEE*, and Tieniu Tan, *Fellow, IEEE*

***Abstract*—The user-generated content (UGC) is a type of dyadic information that provides description of the interaction between users and items (such as rating, purchasing, etc.). Most conventional methods incorporate either a user profile or the item description, which cannot well utilize this kind of content information. Some other works jointly consider user ratings and reviews, but they are based on the factorization technique and have difficulty in providing explanations on generated recommendations. In this study, a coupled topic model (CoTM) for recommendation with UGC is developed. By combining UGC and ratings, the method discussed in this study captures both the content-based preferences and collaborative preferences and, thus, can explain both the user and item latent spaces using the topics discovered from the UGC. The learned topics in CoTM can also serve as proper explanations for the generated recommendations. Experimental results show that the proposed CoTM model yields significant improvements over the compared competitive methods on two typical datasets, that is, MovieLens-10M and Citation-network V1. The topics discovered by CoTM can be used not only to illustrate the topic distributions of users and items, but also to explain the generated user–item recommendations.**

***Index Terms*—Collaborative filtering (CF), recommender systems (RS), topic model, user-generated content (UGC).**

I. Introduction

To HELP users with the problem of information overload [14], [14], [40] when using the Internet, Recommender Systems (RS) manipulate historical behaviors of users and provide users with the most appropriate items. These systems can not only alleviate the effect of information overload and enhance user satisfaction, but also support e-commerce. Generally, most RS can be classified as content-based recommendation [23], collaborative recommendation [36], and hybrid recommendation [9]. In content-based recommendation, one tries to recommend items similar to those the current user liked in the past [1], whereas collaborative techniques exploit the similarity between users and recommended items that similar users liked [4].

Because of the challenging environment of real-world applications, there are still some drawbacks in these techniques. Content-based recommenders have difficulty in finding unexpected items, the serendipity problem, which highlights the tendency of content-based systems in producing recommendations with limited novelty [23]. On the other hand, collaborative filtering (CF) depends on the rating overlap across users and suffers from overfitting on sparse ratings, where similar neighbors are difficult to be detected [38]. Moreover, such systems need not only recommend items, but also convince users to accept (read, buy, listen, or watch) those recommendations [39]. These CF methods always represent users and items with generated latent values and have difficulty in providing interpretability for these values. Besides, these methods just calculate the preference value for a specific user–item pair, but this value lacks convincing explanations that show why the proposals made by the system are reasonable.

Combining content-based and CF, the hybrid approach to RS [9] is usually more effective. Utilizing topic models, several methods [3], [24], [35], [44] can model user profiles or item descriptions for recommendation. Some works, such as TopicMF [5] and GTRT [10], incorporate user ratings and reviews to improve the accuracy of rating prediction. However, both methods are based on the factorization technique and still have the limitations of CF mentioned above. Factorization machine (FM) [29] models the pairwise interaction between variables but cannot capture the latent correlations between the content and the user/item. The model in [2] is also a latent factor model but cannot capture the latent semantic feature of the corpus. These methods [2], [5], [10], [29] have difficulty in providing explanations on the user–item recommendations.

The properties of the user-generated content (UGC) are utilized to alleviate the limitations of conventional approaches mentioned above. Increasingly adopted with the advent of Web 2.0, UGC helps users to collaboratively generate, rather than merely consume content. UGC broadly exists in real-world applications, such as tags allocated by a user to a specific movie (e.g., MovieLens[1]), articles published by an author at a specific conference or journal (e.g., DBLP[2] and Arnetminer[3]), content provided by an editor on a specific Wikipage (e.g., Wikipedia[4]), and so on. In the recommendation scenarios, UGC is assigned

Manuscript received May 22, 2015; revised November 8, 2015, February 8, 2016, and April 18, 2016; accepted May 19, 2016. Date of publication August 3, 2016; date of current version November 11, 2016. This work was supported in part by the National Basic Research Program of China under Grant 2012CB316300 and in part by the National Natural Science Foundation of China under Grant 61403390 and Grant U1435221. This paper was recommended by Associate Editor P. Antunes. *(Shu Wu and Weiyu Guo are co-first authors.)*

S. Wu, Y. Huang, L. Wang, and T. Tan are with the Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100864, China (e-mail: shu.wu@nlpr.ia.ac.cn; yzhuang@nlpr.ia.ac.cn; wangliang@nlpr.ia.ac.cn; tnt@nlpr.ia.ac.cn).

W. Guo is with the School of Engineering Science, University of the Chinese Academy of Sciences, Beijing 100049, China (e-mail: weiyu.guo@ia.ac.cn).

S. Xu is with IBM Research China, Beijing 100193, China (e-mail: xsxuxs@cn.ibm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/THMS.2016.2586480

[1] http://www.movielens.org/
[2] http://www.dblp.org/
[3] http://www.arnetminer.org/
[4] http://www.wikipedia.org/

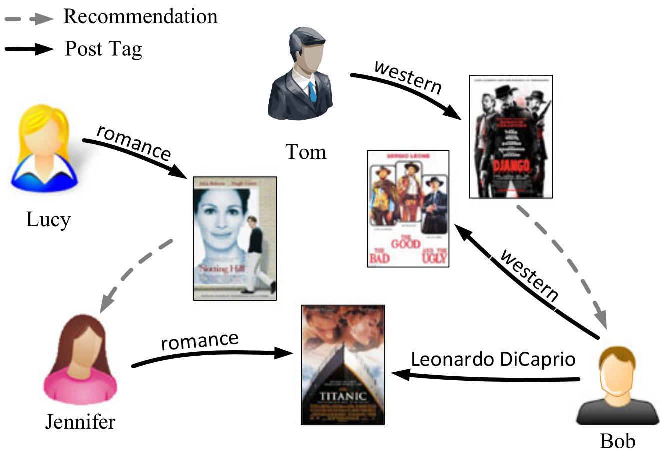


Fig. 1. Illustration of interaction-wise UGC in a movie recommendation system. The solid lines represent "Tagging," and the dashed lines denote "Recommendation." For example, Jennifer tags the movie "Titanic" with "romance" and is recommended to watch "Notting Hill."

by certain users to certain items in the process of user–item interaction. This kind of UGC is named as interaction-wise UGC, that is, content information associated with user–item interaction. Recommendation with interaction-wise UGC includes movie/image recommendation based on the user-generated tags and conference/journal recommendation based on user-published articles.

UGC can provide insights about user preferences and item characteristics [13], [42]. By incorporating the UGC, recommendation systems have the potential to reflect more subtle information, which may not be revealed by very sparse rating behavior, and can support explanations for user preferences, item characteristics, and recommendations. In this study, movie tags are taken as a specific example of UGC. As shown in Fig. 1, both Jennifer and Bob have watched the movie "Titanic." From this behavior, it can be inferred that they have the same preference. However, through further observation of their UGC on "Titanic," it can be realized that Jennifer prefers "romantic" movies, whereas Bob is a fan of "Leonardo DiCaprio." These watching behaviors cannot differentiate users' preferences, although the discriminative UGC can reflect more subtle information. Utilizing the UGC, thus, supports addressing the rating sparsity problem. Second, as the UGC explicitly captures user preferences and item characteristics, RS can demonstrate the topics of user preferences and the topics of item characteristics, instead of extraneous details in latent vectors of conventional methods. Extending the example, in Fig. 1, "Django Unchained" can be recommended to Bob with a proper explanation such as, "it is a western movie starring Leonardo DiCaprio." This explanation may make the recommendation more acceptable than predicted preference values of conventional methods. Besides, the UGC can describe the items more precisely and accurately owing to the "wisdom of the crowd." When judgments are made by a group of people, the aggregated judgment might be better than the best person in the group [47].

In this paper, a new model named coupled topic model (CoTM) is proposed, integrating the UGC with a rating matrix to alleviate the limitations of conventional approaches mentioned above. CoTM aims to learn user preferences and item characteristics from both ratings and UGC to make more accurate recommendations with persuasive explanations. The learned user preferences in CoTM reflect both the content-based preference, representing currently observed and apparent interests, and collaborative preferences, revealing the potential but not-yet-discovered interests in new domains. The proposed model can exhibit user factors and item factors in an explicit manner, which can be interpreted as a vector of topics discovered from UGC. Moreover, the learned topics provide a proper explanation of the recommendation proposals.

The contributions are as follows.

1) A novel latent factor model is proposed, incorporating the UGC into the traditional recommendation systems. By adopting the topic-level vector and the factor-level vector, CoTM is capable of modeling UGC and ratings jointly.
2) The method in this paper captures both collaborative and content-based preferences. The content information can also alleviate the overfitting problem caused by sparse rating data.
3) CoTM can exhibit user preferences and item characteristics in an explicit manner, which can be interpreted as a vector of topics discovered from UGC. The learned topics provide a proper explanation of the proposals made by the system.
4) An efficient approximate inference algorithm based on variational expectation-maximization (EM) methods to train the model is also developed; this algorithm does not require any parameter tuning.

The rest of this paper is organized as follows. Section II reviews related work on content-based filtering, CF, and hybrid recommendation with UGC. Section III shows relevant works, which motivate this study. Section IV introduces the problem statement, document construction, and the proposed CoTM. A detailed variational EM algorithm to estimate the parameters of the CoTM model is developed in Section V. Experiments on two real-world datasets are discussed in Section VI. Finally, the conclusion and future directions of the research are presented in Section VII.

II. Related Work

A. Content-Based Filtering

Content-based filtering [11], [23], [26] is a classic approach to RS, which analyzes a set of descriptions of items previously rated by a user and constructs the profile of user preference based on the features of these items. In other words, this approach tries to recommend items that are similar to those that a user liked in the past. Foltz and Dumais [11] present a content-based information filtering system, matching user preferences to text documents using two methods and two types of user profiles. The LIBRA system is a book recommender that uses a Bayesian learning algorithm and extracts information from books for text categorization [26]. Lops, de Gemmis, and Semeraro [23] reviews the field of content-based recommendation, including a method for representing items and user profiles, and a method for comparing items with the user to determine what to recommend. Content-based filtering is the best approach when content information for users and items is easy to obtain [26]. However, it suffers from limitations such as serendipitous recommendations and quality assessment of filtering items.

B. Collaborative Filtering

CF [1], [36] collects and merges user preference information and generates predictions for an individual user based on the similarity measurements of users and items. Neighborhood-based methods predict ratings based on a matrix of similarity values between items [33] or alternatively, between users [12]. For instance, the item-based method used in [33] models the preference of a user–item pair based on ratings of similar items assigned by the same user.

As a model-based approach, matrix factorization (MF) [7], [19] maps both users and items to a common latent factor space $\mathbb{R}^K$ (each dimension represents a latent factor). Each user i is associated with a factor vector $\eta_{U,i}$, which represents the user preference, and each item j is associated with a factor vector $\eta_{V,j}$, representing the item characteristic. For a specific user–item pair (i, j), the rating value $R_{i,j}$ is predicted as the inner product of latent factor vectors $(\eta_{U,i}^T \eta_{V,j})$. The major point of MF is how to learn user and item latent vectors. One simple approach to perform MF is the singular value decomposition (SVD), which learns a low-rank approximation of rating matrix R. However, the conventional SVD is highly prone to overfitting. Various extensions of SVD have been proposed for alleviating overfitting to some extent, such as regularized MF [45], [48], nonnegative matrix factorization (NMF) [34], and max-margin MF [30], [46]. Several probabilistic interpretations of MF have also been proposed. Salakhutdinov and Mnih present a probabilistic approach to MF named probabilistic matrix factorization (PMF) [31] and then propose a fully Bayesian extension PMF (BPMF) [32] in which the Gibbs algorithm is adopted for inference. Variational approximation methods [18], [43] are also applied to PMF [22], [27].

C. Hybrid Recommendation With User-Generated Content

The hybrid approach to RS [9] typically combines content-based and CF. The hybrid recommendation can be implemented in several ways [1]. For example, by adding content-based characteristics to a collaborative-based method (or vice versa) [4]; or by combining predictions obtained separately using a content-based method and a CF method [25]; or by model unification [6], [28].

In contrast with the attributes, which are descriptions of items in conventional content-based filtering and hybrid recommendation, UGC is a special kind of information depicting the interactions between users and items. Topic models, such as probabilistic latent semantic analysis [15] and latent Dirichlet allocation (LDA) [8], have been implemented for UGC. Utilizing topic models for recommendation, several methods, such as in [3], [24], [35], and [44], incorporate the side information of user profiles or item descriptions. Some works consider combining user ratings and reviews for improving the accuracy of rating prediction. TopicMF [5] uses an MF technique to factorize rating matrix and review text and relates these two tasks by designing the transform function. GTRT [10] explores the review content via a latent factor model and proposes two strategies to leverage the review content as a guidance and a regularization term. These methods are both based on the factorization technique. FM [29] is a general framework that captures the pairwise interaction between variables, but cannot capture the latent correlations between the content and the user/item. The model in [2] is also a latent factor model based on a multiplicative function of row and column latent factors, which are estimated through separate regressions on known row and column features, but cannot capture the latent semantic feature of the corpus. Both methods have difficulty in providing explanations on the user–item recommendations.

Tag recommendation is a UGC recommendation task, which predicts a personalized tag list for a given user–item pair. Several techniques have been used for tag recommendation, such as tensor factorization [37], topic models [20], and graph-based methods [16]. Tso-Sutter, Marinho, and Schmidt-Thieme [41] extend the neighborhood-based recommendation by incorporating tags and ratings for similarity computation. Szomszor *et al.* [38] build tag-clouds for each user and item and subsequently predict the unobserved ratings by comparing the corresponding tag-clouds. Despite this tag recommendation research, these methods cannot be directly implemented for the specific problem of integrating UGC and ratings for item recommendation.

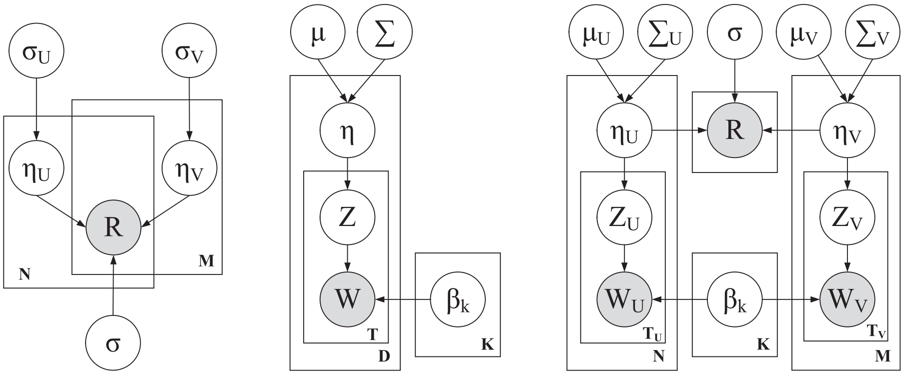
III. Background

This section introduces prior works on PMF [31] and correlated topic model (CTM) [21], which has inspired the construction of the model in this study. PMF is a classical MF framework utilizing the rating matrix for recommendation, while CTM provides a good way of incorporating UGC.

A. Probabilistic Matrix Factorization

PMF [31] is a probabilistic approach to MF; the graphical model representation is illustrated in Fig. 2(a). In a scenario of rating systems, suppose we have N users and M items. R is the incomplete rating matrix, with the element $R_{i,j}$ representing the rating value of user i provided to item j. $\eta_{U,i}$ and $\eta_{V,j}$ are used to represent K-dimensional latent factor vectors of user i and item j, respectively. The conditional distribution over the observed ratings $R \in \mathbb{R}^{N\times M}$ and the prior distributions over $\eta_U \in \mathbb{R}^{K\times N}$ and $\eta_V \in \mathbb{R}^{K\times M}$ can be written as follows:

$$\begin{aligned} p\left(R|\eta_U, \eta_V\right) &= \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\mathcal{N}\left(\eta_{U,i}^T \eta_{V,j}, \sigma^2\right)\right]^{I_{i,j}} \\ p\left(\eta_U|\sigma_U^2\right) &= \prod_{i=1}^{N}\mathcal{N}(\eta_{U,i}|0, \sigma_U^2\mathbf{I}) \\ p\left(\eta_V|\sigma_V^2\right) &= \prod_{j=1}^{M}\mathcal{N}(\eta_{V,j}|0, \sigma_V^2\mathbf{I}) \end{aligned} \tag{1}$$



(a) (b) (c)

Fig. 2. Graphical models of PMF, CTM, and CoTM. (a) There are N users and M items. R is the rating matrix. (b) There are D documents in the corpus. (c) In CoTM, there is a document for each user or item. Therefore, there are N documents for all the users and M documents for all the items. Z_U and $\eta_{U,i}$ are the latent topics and K-dimensional latent factor vectors of user i. β_k is a distribution over the vocabulary.

where $I_{i,j}$ is the indicator function. When user i rated item j, then $I_{i,j} = 1$; otherwise, $I_{i,j} = 0$. The factor vectors η_U and η_V are estimated through maximizing the log-posterior with fixed hyperparameters σ, σ_U, and σ_V in the following equation:

$$\ln p\left(\eta_U, \eta_V | R, \sigma^2, \sigma_U^2, \sigma_V^2\right) = \ln p\left(R | \eta_U, \eta_V, \sigma^2\right) + \ln p\left(\eta_U | \sigma_U^2\right) + \ln p\left(\eta_V | \sigma_V^2\right) + C \quad (2)$$

where C is a constant that does not depend on η_U and η_V. Maximizing this posterior distribution with respect to η_U and η_V with fixed hyperparameters is equivalent to minimizing the sum-of-squares error function with quadratic regularization terms

$$E = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{i,j}\left(R_{i,j} - \eta_{U,i}^T \eta_{V,j}\right)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{N}\|\eta_{U,i}\|_2^2 + \frac{\lambda_V}{2}\sum_{j=1}^{M}\|\eta_{V,j}\|_2^2 \quad (3)$$

where $\lambda_U = \sigma^2/\sigma_U^2$ and $\lambda_V = \sigma^2/\sigma_V^2$.

B. Correlated Topic Model

Topic models have become popular for learning low-dimensional representations of documents. In topic models, each document is associated with a K-dimensional topic vector θ (also called topic proportions), and each topic β is modeled as a distribution over a fixed vocabulary. Instead of being drawn from a Dirichlet distribution such as LDA [8], the topic vector θ in CTM [21] is generated from a logistic transformation of latent factor vector η, which is sampled from Gaussian distribution. Let $W_{d,t}$ be the tth word in document d. The generative process for $W_{d,t}$ is as follows.

1) Draw factor vector $\eta_d \sim \mathcal{N}(\mu, \Sigma)$.
2) For each word $W_{d,t}$
   a) draw topic assignment $z_{d,t}|\eta_d \sim \textit{Mult}\pi(\eta_d))$;
   b) draw word $W_{d,t}|z_{d,t} \sim \textit{Mult}(\beta_{z_{d,t}})$.

Here, $\pi(\eta_d)$ is the logistic transformation function mapping factor vector η_d to topic vector θ_d (within the range of $[0, 1]$)

$$\theta_d = \pi(\eta_d) = \frac{\exp\{\eta_d\}}{\sum_k \exp\{\eta_{d,k}\}}. \quad (4)$$

This process is illustrated as a probabilistic graphical model in Fig. 2(b). CTM can capture topic correlations because of the use of logistic normal distribution.

We intend to exploit the UGC with the topic model. However, a K-dimensional Dirichlet random variable θ is restricted in the $(K-1)$-simplex (a K-dimensional vector θ lies in the $(K-1)$-simplex if $\theta_k \geqslant 0, \sum_{k=1}^{K}\theta_k = 1$), making it not flexible enough to be regarded as factor vectors for the rating prediction task. In CTM, as θ_d is normalized from η_d, θ_d can be seen as a normalized representation of document d, and η_d as an unnormalized representation. The logistic normal distribution provides us a way of combining the PMF and the topic model together.

IV. Coupled Topic Model

In this section, the CoTM is presented, which is a probabilistic framework incorporating the rating matrix and UGC seamlessly. The following is the structure of this section: Section IV-A—the formulation of the problem statement and notations; Section IV-B—the transformation UGC into user and item documents; Section IV-C—the construction of model based on these documents and observed ratings; and Section IV-D—a regularization-based interpretation of CoTM is provided.

A. Problem Statement

To state the problem clearly, the notations in Table I is consistently used to describe the variables in model construction and parameter estimation. Suppose we have N users and M items. Let $R_{i,j}$ denote the rating of user i for item j and $W_{i,j}$ denote the bag-of-words representation of UGC $d_{i,j}$. A novel model is proposed to predict the rating of the unobserved user–item pair (i, j) and the corresponding explanation utilizing the given ratings and the UGC.

TABLE I
SUMMARY OF NOTATIONS

Notation	Description
$d_{i,j}$	UGC assigned by user i to item j
$d_{U,i}$	user document specific to user i
$d_{V,j}$	item document specific to item j
$W_{i,j}$	bag-of-words representation of UGC $d_{i,j}$
$W_{U,i,t}$	the tth word in document $d_{U,i}$
$W_{V,j,t}$	the tth word in document $d_{V,j}$
n_i	word frequency vector of document $d_{U,i}$
n_j	word frequency vector of document $d_{V,j}$
$T_{i,j}$	word count of document $d_{i,j}$
$T_{U,i}$	word count of $d_{U,i}$, $T_{U,i} = \sum_{v=1}^{V} n_{i,v}$
$T_{V,j}$	word count of $d_{V,j}$, $T_{V,j} = \sum_{v=1}^{V} n_{j,v}$
$I_{i,j}$	if user i rates item j, $I_{i,j} = 1$, else $I_{i,j} = 0$
$R_{i,j}$	rating by user i to item j
$\theta_{U,i}$	topic vector of user i
$\theta_{V,j}$	topic vector of item j
$\eta_{U,i}$	factor vector of user i
$\eta_{V,j}$	factor vector of item j
$Z_{U,i,t}$	topic assignment of $W_{U,i,t}$
$Z_{V,j,t}$	topic assignment of $W_{V,j,t}$
β_k	topic k. A topic β_k is a distribution over the vocabulary, a point on the $V - 1$ simplex.

B. Document Extraction

In this section, a simple but effective strategy of extracting user document $d_{U,i}$ and item document $d_{V,j}$, respectively, from the UGC $d_{i,j}$ is introduced.

All the UGC related to user i are simply treated as document $d_{U,i}$, and all the UGC associated with item j as document $d_{V,j}$. Therefore, each user or item has one document. These documents serve as the word-level description of user interests or item characteristics. To reflect a user's changing tastes or item upgrading over time, the document extraction procedure provides RS a way to update or revise the user interest and item characteristic by modifying words in the corresponding documents. Another advantage of this transformation is that user documents and item documents can share the same V-term vocabulary, which makes it possible to embed users and items into the same latent semantic space β and compare them directly. As a consequence, user and item factors with semantic topics can be more easily understood than the latent vectors of conventional MF approaches.

Let $tf(w, d)$ be the frequency of word w in document d. A document d is represented by a list of words. Thus, the vector of word frequency in document d is $n = [tf(w = 1, d), tf(w = 2, d), \ldots, tf(w = V, d)]$. In a user document $d_{U,i}$, the word frequency can be formulated as $tf(w, d_{U,i}) = \sum_{j=1}^{M} tf(w, d_{i,j})$. Similarly, in an item document $d_{V,j}$, the word frequency can be computed as $tf(w, d_{V,j}) = \sum_{i=1}^{N} tf(w, d_{i,j})$. If there is no UGC content of the interaction of user i and item j, we set all the word frequency in document $d_{i,j}$ as zero.

C. Generative Process of the Coupled Topic Model

After document construction, the proposed model is constructed based on all these documents and observed ratings. Each user or item is associated with a K-dimensional topic-level representation θ and a K-dimensional factor-level representation η. The topic vector θ is obtained by taking the logistic transformation of factor vector η to ensure $\sum_{k=1}^{K} \theta_k = 1$. There is also a topic β_k corresponding to the kth dimension of η and θ.

In the training process, representations of the users and items are estimated not only from the documents W_U and W_V, but also from observed ratings R. The generative process for all the user documents W_U, item documents W_V, and observed ratings R is as follows.

1) For each user i:
   a) draw user factor vector $\eta_{U,i} \sim \mathcal{N}(\mu_{\mathrm{U}}, \Sigma_U)$;
   b) for each word $W_{U,i,t}$
      i) draw topic assignment $Z_{U,i,t} | \eta_{U,i} \sim$ *Mult*$(\pi(\eta_{U,i}))$;
      ii) draw word $W_{U,i,t} | Z_{U,i,t} \sim$ *Mult* $(\phi_{Z_{U,i,t}})$.
2) For each item j:
   a) draw item factor vector $\eta_{V,j} \sim \mathcal{N}(\mu_{\mathrm{V}}, \Sigma_V)$;
   b) for each word $W_{V,j,t}$
      i) draw topic assignment $Z_{V,j,t} | \eta_{V,j} \sim$ *Mult*$(\pi(\eta_{V,j}))$;
      ii) draw word $W_{V,j,t} | Z_{V,j,t} \sim$ *Mult*$(\phi_{Z_{V,j,t}})$.
3) For each user–item pair (i, j), draw the rating $R_{i,j} \sim \mathcal{N}(\eta_{U,i}^T \eta_{V,j}, \sigma^2)$.

Here, $\pi(\eta)$ is a logistic transformation function, which has been defined in (4). The graphical model for this generative process is depicted in Fig. 2(c). CoTM maps both users and items to a shared latent space that can be explained by topics β. The factor vectors η are constrained by both the ratings and documents, enabling CoTM prevent overfitting in the training phase. Here, R and β work like two bridges coupling two topic models together. The key issue making this coupling seamless is the adoption of topic-level vectors and factor-level vectors. Owing to the use of logistic normal distribution, CoTM can learn a topic vector and a factor vector for each user and item.

For the ease of exposition, let $\Theta = [\mu_U, \mu_V, \Sigma_U, \Sigma_V, \beta, \sigma]$ denote the model parameters, $\Delta = [\eta_U, \eta_V, Z_U, Z_U]$ denote the latent variables. The joint probability distribution for the observed variables can be written as

$$p(R, W_U, W_V) = \iint p(\eta_U | \sigma_U) p(\eta_V | \sigma_V) p(R | \eta_U, \eta_V) \\ p(W_U | \eta_U) p(W_V | \eta_V) d\eta_U d\eta_V. \quad (5)$$

The model parameters Θ is omitted for brevity. Note that, if we do not take account of the UGC and set $\mu_U = \mu_V = \mathbf{0}$, $\Sigma_U = \sigma_U^2 \mathbf{I}$ and $\Sigma_V = \sigma_V^2 \mathbf{I}$, this special case of CoTM equals PMF.

D. New Perspective of User-Generated Content Constraints

The method of maximum *a posteriori* estimation can be used to obtain a point estimate of the posterior based on observed data [22]. The log of the posterior distribution over the user and item factors is given by

$$\log p(\eta_U, \eta_V | R, W_U, W_V, \Theta) = \log p(\eta_U | \mu_U, \Sigma_U) \\ + \log p(\eta_V | \mu_V, \Sigma_V) + \log p(R | \eta_U, \eta_V) \\ + \log p(W_U | \eta_U, \beta) + \log p(W_V | \eta_V, \beta) + C \quad (6)$$

where C is a constant that does not depend on the latent factors η_U and η_V. If Gaussian parameters $\mu_U = \mu_V = 0$ and $\Sigma_U = \sigma_U^2 \mathbf{I}, \Sigma_V = \sigma_V^2 \mathbf{I}$ are set, the conditional distribution over the observed ratings $R \in \mathbb{R}^{N \times M}$, the prior distributions over $\eta_U \in \mathbb{R}^{K \times N}$ and $\eta_V \in \mathbb{R}^{K \times M}$ can be used to replace the terms, that is, $p(R|\eta_U, \eta_V)$, $p(\eta_U|\mu_U, \Sigma_U)$, and $p(\eta_V|\mu_V, \Sigma_V)$, respectively. Thus, similar to PMF mentioned in Section III, maximizing the log-posterior over the user and item factors with fixed parameter Θ is equivalent to minimizing the sum-of-squared-errors objective function with quadratic regularization terms and the UGC constraints

$$\begin{aligned}\log p(\eta_U, \eta_V|R, W_U, W_V, \Theta) = \\ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{i,j}(R_{i,j} - \eta_{U,i}^T \eta_{V,j})^2 + \frac{\sigma^2}{2\sigma_U^2}\sum_{i=1}^{N}\|\eta_{U,i}\|_2^2 \\ + \frac{\sigma^2}{2\sigma_V^2}\sum_{j=1}^{M}\|\eta_{V,j}\|_2^2 - \sigma^2\sum_{i=1}^{N}\log p(W_{U,i}|\eta_{U,i}, \beta) \\ - \sigma^2\sum_{j=1}^{M}\log p(W_{V,j}|\eta_{V,j}, \beta). \qquad (7)\end{aligned}$$

The parameter σ^2 serves as a confidence parameter, which balances between the MF and the UGC constraint term. If σ^2 is a small value, the model tends to predict the unobserved ratings mainly from the observed ratings R and does not care much about the UGC. Therefore, the MF term would be the major determinants for predicting the ratings, and the UGC counts less. In other cases, if the UGC is trustful and should account for a large proportion in the recommendation task, σ^2 can be set to a larger value. This study adopts a method to learn the parameter rather than set manually.

V. Parameter Estimation

A general technique for finding the maximum likelihood estimators in latent variable models is the EM algorithm. However, the exact posterior distributions of the latent variables Δ are computationally intractable. In this study, variational EM methods are employed for parameter estimation in CoTM.

In the E-step, the latent variables can be inferred by computing the expectation of the posterior distribution $p(\Delta| R, W_U, W_V, \Theta)$. Because of the intractability for exact inference, the variational inference algorithm is used to approximate this posterior. In the M-step, model parameters Θ can be updated given the latent variables Δ. The M-step can be solved by coordinate ascent optimization. Note that, for a compact illustration, only the key mathematical results are presented in the following inference, and the details of the inference procedure are provided in Appendix A.

A. E-step: Variational Inference

The key issue in the E-step is to compute the posterior distribution of latent variables $\Delta = [\eta_U, \eta_V, Z_U, Z_U]$. Given all documents and observed ratings, the posterior distribution of all latent variables $p(\Delta|R, W_U, W_V)$ is intractable to compute. Thus, to approximate this posterior, we appeal to variational methods.

Using Jensen's inequality, the lower bound on the log-likelihood function [18], [43] can be obtained

$$\begin{aligned}\log p(R, W_U, W_V|\Theta) \\ \geqslant E_q[\log p(\Delta, R, W_U, W_V|\Theta)] - E_q[\log q(\Delta)] \quad (8)\end{aligned}$$

where the expectation is taken with respect to q, which is a variational distribution of the latent variables Δ. The right-hand side of the (8) is denoted by $\mathbb{L}$, which is also the lower bound on the log-likelihood function. It can be easily verified that the difference between the log likelihood $\log p(R, W_U, W_V|\Theta)$ and the lower bound $\mathbb{L}$ is the KL divergence between the variational posterior probability and the true posterior probability

$$\log p(R, W_U, W_V|\Theta) = \mathbb{L} + KL(q(\Delta) \parallel p(\Delta|R, W_U, W_V)). \quad (9)$$

The maximum of the lower bound $\mathbb{L}$ occurs when the Kullback–Leibler (K–L) divergence vanishes, which occurs when $q(\Delta)$ equals the posterior distribution $p(\Delta|R, W_U, W_V)$. As a variational distribution, a fully factorized model is used, where all the variables are independently governed by different distributions

$$\begin{aligned} q(\Delta) &= \prod_{i=1}^{N} q(\eta_{U,i}, Z_{U,i}) \prod_{j=1}^{M} q(\eta_{V,j}, Z_{V,j}) \\ q(\eta_{U,i}, Z_{U,i}) &= \prod_{k=1}^{K} q(\eta_{U,i,k}|\lambda_{U,i,k}, \upsilon_{U,i,k}^2) \prod_{t=1}^{T_{U,i}} q(Z_{U,i,t}|\phi_{U,i,t}) \\ q(\eta_{V,j}, Z_{V,j}) &= \prod_{k=1}^{K} q(\eta_{V,j,k}|\lambda_{V,j,k}, \upsilon_{V,j,k}^2) \prod_{t=1}^{T_{V,j}} q(Z_{V,j,t}|\phi_{V,j,t}) \end{aligned} \quad (10)$$

where λ and υ^2 are Gaussian parameters, and ϕ is a K-dimensional multinomial parameter. For ease of explanation, we let $\Omega = [\lambda_U, \lambda_V, \upsilon_U^2, \upsilon_V^2, \phi_U, \phi_V]$ denote the variational parameters. The optimum is achieved at $q(\Delta) = p(\Delta|R, W_U, W_V)$ when the lower bound $\mathbb{L}$ is maximized with respect to the variational parameters. Because of the coupling of η_U and η_V in the term $p(R|\eta_U, \eta_V)$, we adopt alternating optimization by approximating the posterior $p(\eta_U, Z_U)$ with $p(\eta_V)$ kept constant. Subsequently, approximate the posterior $p(\eta_V, Z_V)$ with $p(\eta_U)$ kept constant. That is to say, first λ_U, υ_U, and ϕ_U are optimized given λ_V υ_V; then, λ_V, υ_V, and ϕ_V are optimized given λ_U υ_U. Details of this optimization for CoTM are given in the Appendix, and the variational inference procedure is summarized in Algorithm 1.

B. M-Step: Parameter Estimation

In the M-step, the lower bound $\mathbb{L}$ is maximized with respect to the model parameters $\Theta = [\mu_U, \mu_V, \Sigma_U, \Sigma_V, \beta]$. This amounts to maximum likelihood estimation of the parameters using expected sufficient statistics, where the expectation in $\mathbb{L}$ is taken with respect to the variational distributions computed

Algorithm 1: Variational Inference for CoTM

```
Input: ratings $R$; documents $W_U$, $W_V$; model parameters $\Theta$
Output: variational parameters $\Omega$;
Initialize latent variables $\Delta$ and variational parameters $\Omega$;
for user $i = 1$ to $N$ do
    repeat
        for $t = 1$ to $T_{U,i}$ do
            update $\phi_{U,i,t}$;
        end for
        update $\lambda_{U,i}$ $\upsilon^2_{U,i}$;
    until convergence of $\lambda_{U,i}$ $\upsilon^2_{U,i}$ $\phi_{U,i,t}$
end for
for item $j = 1$ to $M$ do
    repeat
        for $t = 1$ to $T_{V,j}$ do
            update $\phi_{V,j,t}$;
        end for
        update $\lambda_{V,j}$ $\upsilon^2_{V,j}$;
    until convergence of $\lambda_{V,j}$ $\upsilon^2_{V,j}$ $\phi_{V,j,t}$
end for
```

in the E-step

$$\hat{\beta}_k \propto \sum_{i=1}^{N} \phi_{U,i,k} n_i + \sum_{j=1}^{M} \phi_{V,j,k} n_j \tag{11}$$

$$\hat{\mu}_U = \frac{1}{N}\sum_{i=1}^{N} \lambda_{U,i}, \quad \hat{\mu}_V = \frac{1}{M}\sum_{j=1}^{M} \lambda_{V,j} \tag{12}$$

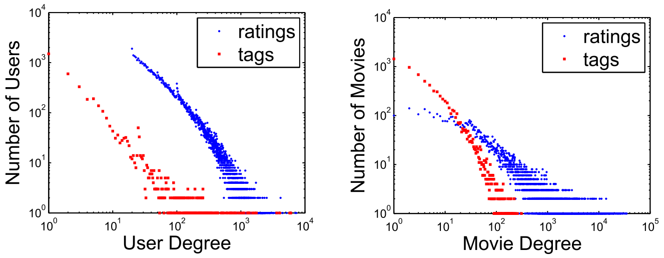
$$\hat{\Sigma}_U = \frac{1}{N}\sum_{i=1}^{N} \mathbf{I}\upsilon^2_{U,i} + (\lambda_{U,i} - \hat{\mu}_U)(\lambda_{U,i} - \hat{\mu}_U)^T \tag{13}$$

$$\hat{\Sigma}_V = \frac{1}{M}\sum_{j=1}^{M} \mathbf{I}\upsilon^2_{V,j} + (\lambda_{V,j} - \hat{\mu}_V)(\lambda_{V,j} - \hat{\mu}_V)^T \tag{14}$$

$$\hat{\sigma}^2 = \frac{1}{R}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{i,j}[(R_{i,j} - \lambda^T_{U,i}\lambda_{V,j})^2 + f(\lambda_{U,i}, \upsilon_{V,j}) + f(\lambda_{V,j}, \upsilon_{U,i}) + f(\upsilon_{U,i}, \upsilon_{V,j})] \tag{15}$$

where n_i is the vector of word counts for document $d_{U,i}$, and n_j is the vector of word counts for document $d_{V,j}$. For any two vectors a, b of the same dimension, $f(a,b) = (a \odot a) \cdot (b \odot b)$, where $\cdot$ is the inner product, and $\odot$ is the Hadamard product, $(a \odot b)_i = a_i b_i$. The E-step and M-step are repeated alternately until the lower bound on the log likelihood converges.

After all the parameters have been estimated, for the (i,j)th entry, the prediction $\hat{R}_{i,j} = \lambda^T_{U,i}\lambda_{V,j}$ can be generated, where $\lambda_{U,i}$ and $\lambda_{V,j}$ are the variational Bayesian estimators of factor vectors η_U and η_V.



(a) (b)

Fig. 3. Degree distributions on the MovieLens-10M dataset. Plots (a) and (b) show degree distributions of user and movie, respectively, on a log–log scale. The user/movie degree on ratings refers to the number of ratings per user/movie, while the user/movie degree on tags refers to the number of tags per user/movie.

VI. Experiments and Analysis

This section conducts experiments on two real-world datasets. Initially, experimental designs and evaluation metrics are discussed. Then, the performance on rating prediction, sparse data, and convergent rate with comparable methods are compared. Finally, the topic discovered by CoTM is illustrated; the topic distributions of users and items are investigated and utilized to explain newly generated user–item interactions.

A. Description of Datasets

1) MovieLens-10M: The MovieLens-10M[5] is a movie rating dataset. In this dataset, each user rated at least 20 movies with the rating values scale of 1–5 and provided tags to movies. Fig. 3 shows basic statistical character about this dataset, that is, the degree on both tags and ratings follows a power law. That is to say, most of the users have a few ratings or tags, whereas a few users have a great many ratings or tags.

In this context, tags are regarded as a kind of UGC. To evaluate the model in this study with UGC, users and movies without tags are removed from the original dataset, and an experimental dataset that contains 323 546 ratings with 1033 users, 1996 movies, and 17 552 tags is constructed. In the procedure of document construction, each user document is composed of all tags posted by the user, and each item document consists of all the tags related to the corresponding movie.

2) Citation-Network V1: Citation-network $V1$[6] is a dataset describing academic articles and their citation relationships. Removing articles that lack integrated publication information from the original dataset, a subset that contains 27 704 authors, 120 637 articles, and 2144 publications is obtained. When an author cites an article appearing in a certain publication, it is argued that this behavior can be treated as a link which is generated between the author and the publication. In this circumstance, the article titles are treated as a kind of UGC between the author and the publication, and the task is to recommend publications to authors. All articles linked to a particular author make up a library that he or she is interested in. Thus, article titles are extracted as a UGC document for the user. Each item (publication) document consists of all corresponding article titles.

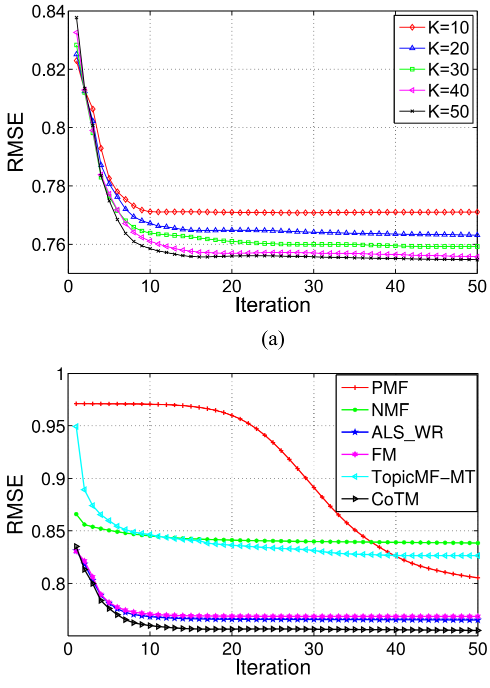
[5] MovieLens is available at: http://www.grouplens.org/node/73

[6] Citation-network V1 is available at: http://arnetminer.org/citation

TABLE II
PERFORMANCE COMPARISON ON THE MOVIELENS AND CITATION-NETWORK DATASETS WITH DIFFERENT FACTOR DIMENSIONALITIES K

	MovieLens-10 M				
Method	$K = 10$	$K = 20$	$K = 30$	$K = 40$	$K = 50$
PMF	0.8113	0.8044	0.8034	0.8001	0.8040
BPMF	0.7921	0.7906	0.7933	0.7990	0.7968
NMF	0.8388	0.8409	0.8337	0.8321	0.8286
ALS-WR	0.7710	0.7683	0.7681	0.7626	0.7625
TopicMF-MT	0.7898	0.8126	0.8251	0.8498	0.8596
FM	0.7780	0.7761	0.7767	0.7766	0.7685
CoTM	**0.7635**	**0.7572**	**0.7560**	**0.7537**	**0.7518**
	Citation-network V1				
Method	$K = 10$	$K = 20$	$K = 30$	$K = 40$	$K = 50$
PMF	0.7540	0.7466	0.7462	0.7468	0.7495
BPMF	0.7528	0.7375	0.7371	0.7336	0.7278
NMF	0.7504	0.7412	0.7342	0.7320	0.7329
ALS-WR	0.7340	0.7265	0.7234	0.7224	0.7138
TopicMF-MT	0.7405	0.7662	0.7827	0.8033	0.8200
FM	0.7091	0.7059	0.7047	0.7052	0.7051
CoTM	**0.6997**	**0.6981**	**0.7069**	**0.7176**	**0.7084**

The predictive accuracy is measured by RMSE.



(a)

(b)

Fig. 4. x-axis shows the number of iterations and the y-axis displays RMSE in test set. (a) indicates convergence rates of CoTM with different dimensionalities. Each line corresponds to a specific dimension. (b) shows the convergence rates of CoTM and compared methods. Each line represents a specific model. The RMSE is the mean of different dimensions ($K = 10, 20, 30, 40, 50$).

Intuitively, the more links between an author and a publication, the more possibility that he/she may be interested in the publication. Therefore, the link numbers are treated as the ratings that indicate user preference values, and a normalization method similar to the normalized techniques mentioned in [17] is adopted to transform this kind of ratings to $[1, 5]$. Specifically, a rating in this context is defined as

$$R_{i,j} = \begin{cases} 5, & \text{if count}(i,j) > 10 \\ \dfrac{4 \times (\text{count}(i,j) - 1)}{9} + 1, & \text{else.} \end{cases}$$

B. Experimental Design and Evaluation Methodology

To examine the performance of the proposed model, CoTM is compared with several state-of-the-art methods such as PMF [31], BPMF [32], NMF [34], alternating-least-squares with weighted-regularization (ALS-WR) [48], FMs [29], and TopicMF-MT [5].

In this experiment, the performance of the methods with various dimensionalities ($K = 10, 20, 30, 40, 50$) are examined, and training and testing on randomly split training (80%) and testing (20%) data are carried out. Moreover, to examine the robustness dealing with sparse rating data, different amounts of ratings (20–80%) are used as training data and the remaining ratings (80–20%) as testing data.

Additionally, as the features of UGC, that is, tags in MovieLens-10M and words in Citation-network V1, usually have numerous dimensions and a lot of noisy terms, the UGC data are preprocessed with the term frequency inverse document frequency (TF-IDF) technique. Specifically, the TF-IDF value is calculated for each term (tag or word) and then sort terms of each document according to their TF-IDF values. Finally, the top-N terms of each document are used to build the dictionaries of words for these two datasets, namely 1490 terms in MovieLens-10M and 10493 terms in Citation-network V1.

1) Parameter Setting: The parameter settings of models are listed here. In the training phase, CoTM is initialized with $\mu_U = \mu_V = \mathbf{0}$, $\Sigma_U = \Sigma_V = \mathbf{I}$, $\sigma = 1$ in all the experiments. The topic β is initialized randomly. For the PMF model, the regularization parameters λ_U and λ_V are set to 0.01. In BPMF, the parameters $\alpha = 2$, $\mu_0 = 0$, and W_0 are set to the identity matrix, for both user and item hyperpriors. The factor vectors in PMF are initialized randomly. The factor vector in BPMF is initialized with the result of PMF as mentioned in [32]. In FM, the user document and the item document are treated as additional features [29]. In TopicMF-MT, the tags of a user–item interaction are treated as the reviews of this user and item pair.

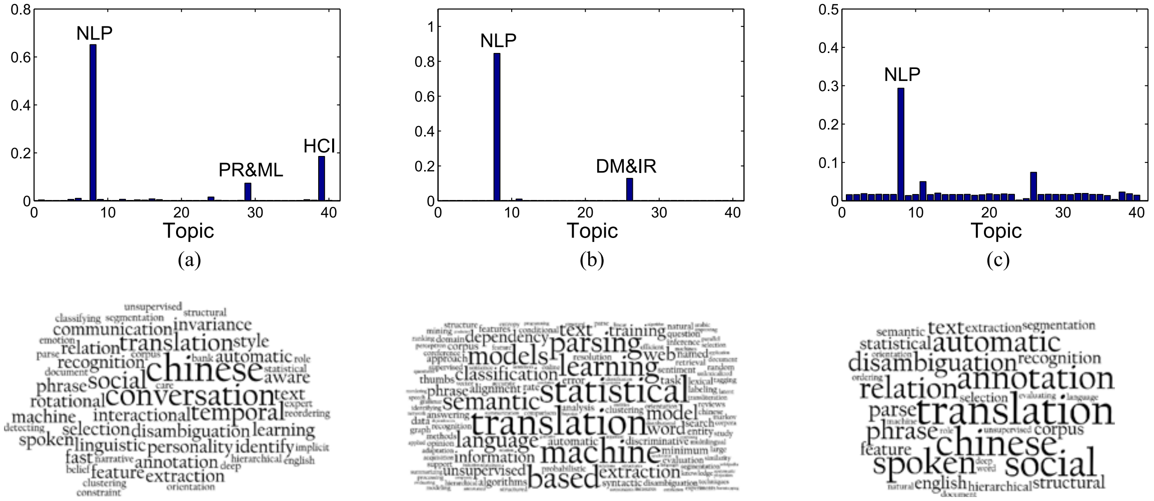
2) Evaluation: The root-mean-square error (RMSE) is adopted to measure the rating prediction performance of the proposed approach in comparison with other latent factor methods. Values close to zero show better performance. RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{(i,j)\in R_t} \left(R_{i,j} - \hat{R}_{i,j}\right)^2}{|R_t|}}$$

where $|R_t|$ denotes the number of test ratings R_t.

C. Experimental Results and Analysis

1) Prediction Performance: Table II gives the performance comparison with a variety of dimensionalities ($K = 10, 20, 30, 40, 50$) on the MovieLens and the Citation-network



(a) (b) (c)

(d) (e) (f)

Fig. 5. Top panels display the topic distribution of Dan Jurafsky, EMNLP, and their interaction, respectively, while the bottom panels show the corresponding word cloud. The word size is proportional to word frequency in the corresponding document. (a) Dan Jurafsky. (b) EMNLP. (c) Interaction of Dan Jurafsky and EMNLP. (d) Dan Jurafsky. (e) EMNLP. (f) Interaction of Dan Jurafsky and EMNLP.

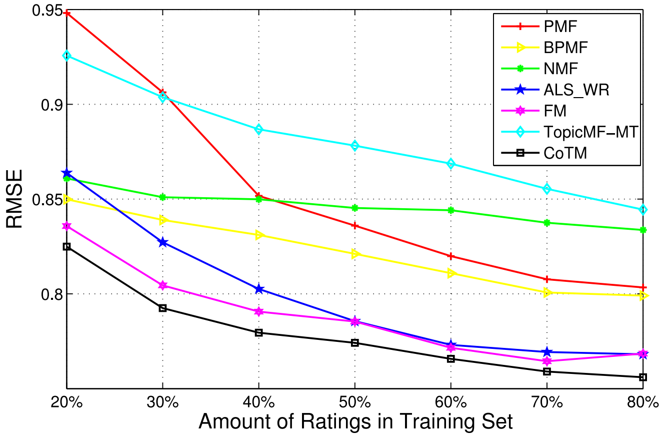


Fig. 6. Performance comparison of different methods on MovieLens with different amount of ratings in training set ($K = 40$).

datasets. On both datasets, comparing with the methods such as PMF, BPMF, NMF, and ALS-WR, which do not consider UGC, the algorithm in this study consistently achieves better performance for every setting of latent dimensionality. Owing much to the UGC, CoTM increases the performance by more than 5% compared with its special case, PMF. Compared with the method ALS-WR, CoTM utilizes the UGC but without weighted regularization that has at least 1% improvement. Interaction-wise UGC offers CoTM greater predictive ability than simply changing the factor dimensionality. Among the methods incorporating UGC, that is, TopicMF-MT, FM, and CoTM, TopicMF-MT achieves the worst performance. Although TopicMF-MT can take reviews into model construction, this model still has difficulty in capturing enough semantic information from each piece of interaction-wise UGC. Although the performance of CoTM may not surpass the performance of FM in some situations, it can still be observed that the best performances on two datasets are obtained by CoTM. In addition, CoTM can exhibit user preferences and item characteristics explicitly using the topics discovered from UGC and provide persuasive explanations of recommendations. This property of CoTM is discussed in Section VI-C4 and Fig. 5.

2) Model Convergence: In this section, the convergence rates of CoTM are first examined with various dimensionalities and then compared with PMF, NMF, ALS-WR, FM, and TopicMF-MT. The performances of convergence on the MovieLens dataset are demonstrated in Fig. 4.

Fig. 4(a) shows the performance of CoTM measured by RMSE in every iteration. Each line illustrates the performance of CoTM with a dimensionality. With the increase of iteration, RMSE values with various dimensionalities decrease greatly at the beginning and then decrease slightly after ten iterations. The content information makes CoTM overcome the problem of overfitting in the training period. In Fig. 4(b), it can be observed that CoTM has great improvement in convergence rates compared with other methods. This figure indicates that the PMF converges very slowly. The performance of NMF has achieved the convergence when the iteration is 10, but performs badly in rating prediction. The ALS-WR and FM can achieve good performance in both rating prediction and model convergence. TopicMF-MT has similar convergence rate with ALS-WR and FM, but obtains poor performance. By incorporating the interaction-wise UGC, CoTM performs better than the baseline methods.

3) Handling Sparse Data: The ability of all methods in handling different amount of training ratings is further evaluated. The performance comparison with $K = 40$ on the MovieLens dataset is shown in Fig. 6. In this part, different amounts of ratings from 20% to 80% are used as the training set to assess all the methods.

The experimental results indicate that CoTM can obtain the best performance with varying amount of training ratings. Especially, when fewer ratings are provided, because of the incorporation of UGC, CoTM outperforms other methods more

TABLE III
SIX TOPICS DISCOVERED BY COTM ON CITATION-NETWORK DATASET: EACH TOPIC IS SHOWN WITH THE FIRST TEN MATCHED WORDS, FIRST SIX MATCHED AUTHORS, AND FIRST SEVEN MATCHED PUBLICATIONS ($K = 40$)

topic 08 “NLP”	topic 14 “Programming”	Topic 15 “Graphics”	topic 17 “Multimedia”	topic 36 “IR”	topic 39 “HCI”
translation	analysis	motion	video	web	mobile
statistical	java	interactive	retrieval	search	iteration
machine	program	render	search	information	user
learning	type	time	information	retrieval	design
semantic	check	animation	content	query	interface
language	static	surface	detection	semantic	social
parsing	verification	data	semantic	text	information
text	test	real	multimedia	mining	context
extraction	detection	simulation	annotation	collaborative	human
word	software	dynamic	web	filtering	display
Chris Callison-Burch	Cormac Flanagan	Mark Pauly	Yu-Gang Jiang	Claudiu S. Firan	Michael Rohs
Philipp Koehn	Mayur Naik	Bart Adams	Cees G. M. Snoek	Shenghua Bao	Daniel Wigdor
Erik Tjong Kim Sang	Koushik Sen	Yizhou Yu	Chong-Wah Ngo	Mingfang Wu	Chia Shen
Alessandro Moschitti	Stephen N. Freund	Alexander Belyaev	Xirong Li	Lichun Yang	Miguel A. Nacenta
Wenliang Chen	Zhong Shao	Tamar Shinar	Lyndon S. Kennedy	Mariam Daoud	Roel Vertegaal
Benjamin Snyder	Patrice Godefroid	Sharif Elcott	Wan-Lei Zhao	Zhicheng Dou	Shahram Izadi
EMNLP	POPL	SIGGRAPH	SIGMM	IPM	NordiCHI
StatMT	SIGPLAN Notices	SCA	CIVR	ICMR	MobileHCI
ACL	OOPSLA	SGP	ICME	IR	IHM
NAACL-HLT	ISSTA	I3D	JVCIR	SIGIR	UAHCI
Machine Translation	TOPLAS	TOG	TOMCCAP	JASIS	OZCHI
EACL	PEPM	SPM	EuroITV	JCDL	IDC
COLING/ACL	PLDI	GDSPM	MTA	CIVR	TEI

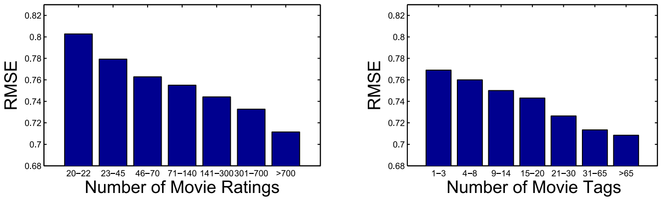


Fig. 7. Collective intelligence analysis: Movies are binned by the number of ratings (left panel) or tags (right panel), with the x-axis showing those bins, and the y-axis displaying the RMSE on the test set for each bin ($K = 40$).

significantly. When 80% of ratings are used as training set, CoTM increases the performance of PMF with 5.9%, while given 20% ratings, CoTM enhances the performance by more than 13%. On the other hand, the regularized variations of MF, that is, NMF, ALS-WR, and TopicMF-MT, outperform PMF because of additional constraints, that is, nonnegative latent factors, weighted regularization, or the regressive term of word frequency matrix.

4) Topic Discovery and Latent Factor Explanation: Another advantage of CoTM is that it can explain the user latent space and item latent space using the topics discovered from the UGC, and these topics can also be utilized to illustrate the reasons of the generated proposals. For user i, the best-matched topics can be found by ranking the entries of factor vector $\eta_{U,i}$. For item j, the entries of factor vector $\eta_{V,j}$ can also be ranked. For a user–item recommendation, the topic distribution of this interaction can be generated by $\pi(\eta_{U,i} \odot \eta_{V,j})$. The best-matched topics can be treated as an explanation of user interest and item characteristic.

To better illustrate the topic discovered by CoTM, the best-matched words, authors, and publications are used to explain each topic discovered from the Citation-network dataset; some of them are displayed in Table III. Specifically, for each topic, the first ten words, first six matched authors, and first seven matched publications are shown. For example, CoTM illustrates the topic 08 as the natural language processing (NLP) with its best-matched words, such as “translation,” “statistical,” and so on, the top authors, such as “Chris C.-B.,” “Philipp K.,” and so on, and the most related publications, such as “EMNLP,” “StatMT,” “ACL,” and so on.

The effect of documents and latent topics are also demonstrated in this study. With the proposed model, the user document d_U makes it possible to present a collection of words users may be interested in. As shown in Fig. 5(d), Dan Jurafsky’s interest is visualized as a word cloud. In real-world applications, it is often difficult to recommend the right item at a right time because of the complex context. The word cloud, obtained from the user document, makes it possible for the user to select the related keyword he/she may like at that time. It will largely diminish the range of searching and make it quicker and easier for users to find what they want. However, the information from the user document is limited and lagging. The learned topics can find the potential keywords a user may be interested in and then enrich the user preference cloud. From the topic distribution in Fig. 5(a), it can be seen that the topic Dan Jurafsky is most interested in is NLP (topic 08). The words in the corresponding topic can be the optional choice of the word cloud.

To explore the reasons behind the recommendation, this study used Dan Jurafsky and EMNLP further as examples. The overlap between the user document and the item document can be a kind of word-level interpretation of user–item interaction. It provides a possible reason why RS should recommend the item to the user. According to the above method, a word cloud for the interaction between Dan Jurafsky and EMNLP can be taken [see Fig. 5(f)]. When RS recommends the publication "EMNLP" to "Dan Jurafsky," the word cloud in Fig. 5(f) serves as a more concise and convincing explanation than the word cloud in Fig. 5(e). However, the above strategy fails when the user document and the item document have few words in common. In this case, the topic-level interpretation can help. The topic distribution of user–item interaction can be generated by $\pi(\eta_{U,i} \odot \eta_{V,j})$. Fig. 5(c) shows the topic distribution of interaction between Dan Jurafsky and EMNLP. From this topic distribution, the user's potential interests about the item can be inferred. To sum up, by combining the word-level interpretation and topic-level interpretation, a topic-enriched word cloud can be considered as an explanation of user–item interaction.

5) Effect of Collective Intelligence: The effect of collective intelligence is also shown. Subjectively, the more times a movie has been tagged or rated, the more precisely this movie would be represented. With this, the movie characteristic can be captured more accurately, thus making better recommendation. To verify the above assumption, the movies are first divided into seven groups according to the number of ratings and the number of tags, and subsequently, prediction accuracies of different groups are evaluated. The experimental result shown in Fig. 7 reveals that more the ratings and the UGC in train data, the better recommendation performance for movies. It suggests that the researchers can pay more attention to the interaction-wise UGC to promote the conventional CF for recommendation.

VII. Conclusion and Discussion

Based on the intuition that UGC might contribute to capturing the user interest and item characteristic, a novel framework of jointly modeling UGC and ratings simultaneously for the recommendation task has been proposed. A variational EM algorithm has also been developed for CoTM, without requiring the tuning of too many parameters. Furthermore, CoTM exhibits an interpretable low-dimensional representation for each user and item using topics discovered from UGC, and these topics can also provide proper explanations for recommendations. Owing to the good properties of interpretation, we suggest that the user experience would be greatly improved in real-world RS. The experimental results on two different datasets, MovieLens-10M and Citation-network V1, have shown that the approach of this study outperforms several CF algorithms. The study indicates that the UGC not only boosts the recommendation performance, but also makes the system more robust in dealing with sparse rating data.

As the content part is the most time consuming in the training procedure, in future extensions, a distributed version of CoTM should be considered. And because the focus was not centered on the document extraction here, a well-defined document can be made in a future study for obtaining a more effective model.

Appendix Details of Variational Inference

A. Variational Objective

The lower bound $\mathbb{L} = T_1 + T_2 + T_3 + T_4 + T_5$, where

$$T_1 = \sum_{i=1}^{N} E_q[\log p(\eta_{U,i}|\mu_U, \Sigma_U)] + \sum_{j=1}^{M} E_q[\log p(\eta_{V,j}|\mu_V, \Sigma_V)]$$

$$T_2 = \sum_{i=1}^{N}\sum_{j=1}^{M} I_{i,j} E_q[\log p(R_{i,j}|\eta_{U,i}, \eta_{V,j})]$$

$$T_3 = \sum_{i=1}^{N} E_q[\log p(Z_{U,i}|\eta_{U,i})] + \sum_{j=1}^{M} E_q[\log p(Z_{V,j}|\eta_{V,j})]$$

$$T_4 = \sum_{i=1}^{N} E_q[\log p(W_{U,i}|Z_{U,i}, \beta)] + \sum_{j=1}^{M} E_q[\log p(W_{V,j}|Z_{V,j}, \beta)]$$

$$T_5 = -\sum_{i=1}^{N} E_q[\log q(\eta_{U,i}, Z_{U,i})] - \sum_{j=1}^{M} E_q[\log q(\eta_{V,j}, Z_{V,j})].$$

Because of the similarity of approximating the posterior $p(\eta_U, Z_U)$ and $p(\eta_V, Z_V)$, we will introduce the inference method just for $p(eta_U, Z_U)$.

In the first term T_1,

$$E_q[\log p(\eta_{U,i}|\mu_U, \Sigma_U)] = \frac{1}{2}\{\log\left|\Sigma_U^{-1}\right| - K\log 2\pi - Tr(diag(\upsilon_{U,i}^2)\Sigma_U^{-1}) + (\lambda_{U,i} - \mu_U)^T \Sigma_U^{-1}(\lambda_{U,i} - \mu_U)\}.$$

In the second term T_2,

$$E_q[\log p(R_{i,j}|\eta_{U,i}, \eta_{V,j})] = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}y_{i,j}^2 + \frac{R_{i,j}}{\sigma^2}\lambda_{U,i}^T\lambda_{V,j} - \frac{1}{2\sigma^2}(\lambda_{U,i}^T\lambda_{V,j})^2 - \frac{1}{2\sigma^2}[f(\lambda_{U,i}, \upsilon_{V,j}) + f(\lambda_{V,j}, \upsilon_{U,i}) + f(\upsilon_{U,i}, \upsilon_{V,j})]$$

where f has been defined in Section V-B.

The third term T_3 is difficult to compute owing to the nonconjugacy of the logistic normal to multinomial. By adopting a new variational parameter $\zeta_{U,i}$ similar to [21], we obtain

$$E_q[\log p(Z_{U,i,t}|\eta_{U,i})] = \sum_{k=1}^{K} \lambda_{U,i,k}\phi_{U,i,t,k} - \zeta_{U,i}^{-1}\left(\sum_{k=1}^{K}\exp\left\{\lambda_{U,i,k} + \upsilon_{U,i,k}^2/2\right\}\right) + 1 - \log\zeta_{U,i.}$$

In the fourth term T_4,

$$E_q[\log p\,(W_{U,i}|Z_{U,i},\beta)] = \sum_{t=1}^{T_{U,i}} \sum_{k=1}^{K} \phi_{U,i,t,k} \log \beta_{k,W_{U,i,t}}.$$

The fifth term T_5 is the entropy of the variational distribution $q\,(\Delta)$,

$$E_q[\log q\,(\eta_{U,i}, Z_{U,i})] = -\frac{K}{2} - \frac{K}{2}\log 2\pi - \sum_{k=1}^{K} \frac{1}{2} \log \upsilon^2_{U,i,k} + \sum_{t=1}^{T_{U,i}} \sum_{k=1}^{K} \phi_{U,i,t,k} \log \phi_{U,i,t,k}.$$

B. Coordinate Ascent Optimization

Finally, we maximize the bound in (8) with respect to the variational parameters $\Omega = [\lambda_U, \lambda_V, \upsilon^2_U, \upsilon^2_V, \phi_U, \phi_V]$. We use a coordinate ascent algorithm, iteratively maximizing the bound with respect to each parameter.

First, we maximize (8) with respect to $\zeta_{U,i}$, $\phi_{U,i,t,k}$, and $\lambda_{U,i,k}$, respectively:

$$\hat{\zeta}_{U,i} = \sum_{k=1}^{K} \exp\{\lambda_{U,i,k} + \upsilon^2_{U,i,k}/2\}$$

$$\phi_{U,i,t,k} \propto \exp(\lambda_{U,i,k})\beta_{k,w_{U,i,t}}.$$

We use the conjugate gradient algorithm with the derivative

$$\frac{dL}{d\lambda_{U,i}} = -\Sigma_U^{-1}(\lambda_{U,i} - \mu_U) - \frac{1}{\sigma^2}\sum_{j=1}^{M} I_{i,j}(\lambda^T_{U,i}\lambda_{V,j}\lambda_{V,j} + \lambda_{U,i} \odot \upsilon^2_{V,j} - R_{i,j}\lambda_{V,j}) + \sum_{t=1}^{T_{U,i}} \phi_{U,i,t,1:K} - \frac{T_{U,i}}{\zeta_{U,i}}\exp\{\lambda_{U,i} + \frac{\upsilon^2_{U,i}}{2}\}.$$

Finally, we maximize with respect to $\upsilon^2_{U,i}$. There is also no analytic solution. We use Newton's method for each coordinate with the constraint that $\upsilon_{U,i} > 0$,

$$\frac{dL}{d\upsilon^2_{U,i,k}} = -\frac{\Sigma^{-1}_{U,k,k}}{2} - \frac{1}{2\sigma^2}\sum_{j=1}^{M} I_{i,j}(\lambda^2_{V,j,k} + \upsilon^2_{V,j,k}) - \frac{T_{U,i}}{2\zeta_{U,i}} \exp\left(\lambda_{U,i,k} + \frac{\upsilon^2_{U,i,k}}{2}\right) + \frac{1}{2\upsilon^2_{U,i,k}}.$$

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which helped to improve the quality of this paper.

REFERENCES

[1] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[2] D. Agarwal and B.-C. Chen, "Regression-based latent factor models," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 19–28.

[3] D. Agarwal and B.-C. Chen, "FLDA: Matrix factorization through latent Dirichlet allocation," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 91–100.

[4] M. Balabanović and Y. Shoham, "FAB: Content-based, collaborative recommendation," *Commun. ACM*, vol. 40, no. 3, pp. 66–72, Mar. 1997.

[5] Y. Bao, H. Fang, and J. Zhang, "TopicMF: Simultaneously exploiting ratings and reviews for recommendation," in *Proc. 28th AAAI Conf. Artif. Intell.*, 2014, pp. 2–8.

[6] J. Basilico and T. Hofmann, "Unifying collaborative and content-based filtering," in *Proc. 21st Int. Conf. Mach. Learn.*, New York, NY, USA, Jul. 2004, p. 9.

[7] R. M. Bell and Y. Koren, "Lessons from the netflix prize challenge," *ACM SIGKDD Explorations Newslett.*, vol. 9, no. 2, pp. 75–79, 2007.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[9] R. Burke, *Hybrid Web Recommender Systems*. Berlin, Germany: Springer, Jan. 2007.

[10] X. Chen, Y. Yao, F. Xu, and J. Lu, "Exploring review content for recommendation via latent factor model," in *Proc. 13th Pacific Rim Int. Conf. Artif. Intell.*, 2014, pp. 668–679.

[11] P. W. Foltz and S. T. Dumais, "Personalized information delivery: An analysis of information filtering methods," *Commun. ACM*, vol. 35, no. 12, pp. 51–60, Dec. 1992.

[12] K. Goldberg, T. Roeder, D. Gupta, and C. Perkins, "Eigentaste: A constant time collaborative filtering algorithm," *Inf. Retrieval*, vol. 4, no. 2, pp. 133–151, Jul. 2001.

[13] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *J. Inf. Sci.*, vol. 32, no. 2, pp. 198–208, 2006.

[14] B. M. Gross, *The Managing of Organizations: The Administrative Struggle*, vol. 2. New York, NY, USA: Free Press of Glencoe, 1964.

[15] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 15th Conf. Uncertainty Artif. Intell.*, 1999, pp. 289–296.

[16] A. Hotho, R. Jäschke, C. Schmitz, G. Stumme, and K.-D. Althoff, "Folkrank: A ranking algorithm for folksonomies," in *Proc. LWA Conf.*, 2006, vol. 1, pp. 111–114.

[17] G. Jawaheer, M. Szomszor, and P. Kostkova, "Comparison of implicit and explicit feedback from an online music recommendation service," in *Proc. 1st Int. Workshop Inf. Heterogeneity Fusion Recommender Syst.*, 2010, pp. 47–51.

[18] M. I. J. Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[19] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[20] R. Krestel, P. Fankhauser, and W. Nejdl, "Latent Dirichlet allocation for tag recommendation," in *Proc. 3rd ACM Conf. Recommender Syst.*, 2009, pp. 61–68.

[21] J. D. Lafferty and D. M. Blei, "Correlated topic models," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2005, pp. 147–154.

[22] Y. J. Lim and Y. W. Teh, "Variational Bayesian approach to movie rating prediction," presented at the KDD Cup Workshop, San Jose, CA, USA, 2007.

[23] P. Lops, M. de Gemmis, and G. Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*. New York, NY, USA: Springer, 2011, pp. 73–105.

[24] J. McAuley and J. Leskovec, "Hidden factors and hidden topics: Understanding rating dimensions with review text," in *Proc. 7th ACM Conf. Recommender Syst.*, New York, NY, USA, 2013, pp. 165–172.

[25] T. Miranda *et al.*, "Combining content-based and collaborative filters in an online newspaper," in *Proc. ACM SIGIR Workshop Recommender Syst.*, 1999.

[26] R. J. Mooney and L. Roy, "Content-based book recommending using learning for text categorization," in *Proc. 5th ACM Conf. Digital Libraries*, New York, NY, USA, Jun. 2000, pp. 195–204.

[27] S. Nakajima and M. Sugiyama, "Implicit regularization in variational Bayesian matrix factorization," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 815–822.

[28] A. Popescul, L. H. Ungar, D. M. Pennock, and S. Lawrence, "Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, 2001, pp. 437–444.

[29] S. Rendle, "Factorization machines with libFM," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, p. 57, 2012.

[30] J. D. M. Rennie and N. Srebro,"Fast maximum margin matrix factorization for collaborative prediction," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 713–719.

[31] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2007, pp. 1257–1264.

[32] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using Markov chain Monte Carlo," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 880–887.

[33] B. Sarwar, G. Karypis, J. Konstan, and J. Reidl, "Item-based collaborative filtering recommendation algorithms," in *Proc. 10th Int. Conf. World Wide Web*, New York, NY, USA, Apr. 2001, pp. 285–295.

[34] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2001, vol. 13, pp. 556–562.

[35] H. Shan and A. Banerjee, "Generalized probabilistic matrix factorizations for collaborative filtering," in *Proc. IEEE 10th Int. Conf. Data Mining*, 2010, pp. 1025–1030.

[36] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. Artif. Intell.*, vol. 2009, p. 4, 2009.

[37] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "Tag recommendations based on tensor dimensionality reduction," in *Proc. ACM Conf. Recommender Syst.*, 2008, pp. 43–50.

[38] M. Szomszor *et al.*, "Folksonomies, the semantic web, and movie recommendation," presented at the 4th Eur. Semantic Web Conf., Innsbruck, Austria, 2007.

[39] N. Tintarev and J. Masthoff, "Designing and evaluating explanations for recommender systems," in *Recommender Systems Handbook*. New York, NY, USA: Springer, 2011, pp. 479–510.

[40] A. Toffler, *Future Shock*. New York, NY, USA: Bantam, 1990.

[41] K. H. L. Tso-Sutter, L. B. Marinho, and L. Schmidt-Thieme, "Tag-aware recommender systems by fusion of collaborative filtering algorithms," in *Proc. ACM Symp. Appl. Comput.*, 2008 pp. 1995–1999.

[42] J. Vig, S. Sen, and J. Riedl, "Tagsplanations: Explaining recommendations using tags," in *Proc. 14th Int. Conf. Intell. User Interfaces*, 2009, pp. 47–56.

[43] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, nos. 1/2, pp. 1–305, 2008.

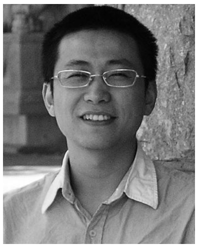
[44] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 448–456.

[45] B. Webb, *Netflix update: Try this at home*, (2006). http://www.sifter.org/∼simon/journal/20061211.html

[46] M. Xu, J. Zhu, and B. Zhang, "Nonparametric max-margin matrix factorization for collaborative prediction," in *Proc. Adv. Neural Inf. Process. Syst. Conf.*, 2012, pp. 64–72.

[47] S. K. M. Yi, M. Steyvers, M. D. Lee, and M. J. Dry, "The wisdom of the crowd in combinatorial problems," *Cognitive Sci.*, vol. 36, no. 3, pp. 452–470, 2012.

[48] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," in *Algorithmic Aspects in Information and Management*. New York, NY, USA: Springer, 2008, pp. 337–348.

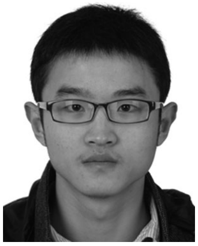


Shu Wu (M'13) received the B.E. degree from Hunan University, Changsha, China, in 2004, the M.E. degree from Xiamen University, Xiamen, China, in 2007, and the Ph.D. degree from the University of Sherbrooke, Sherbrooke, QC, Canada, in 2012, all in computer science.

He is currently an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include data mining and recommendation systems.

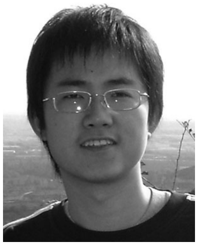
Weiyu Guo received the B.E. degree in computer science and technology from the School of Computer and Information Technology, Shanxi University, Taiyuan, China, in 2010, and the M.E. degree in software engineering from the Software Institute, Nanjing University, Nanjing, China, in 2012. He is currently working toward the Ph.D. degree in the College of Engineering and Information Technology, University of the Chinese Academy of Sciences, Beijing, China.

His research interests include machine learning, recommendation systems, and large-scale Web data mining.



Song Xu received the bachelor's degree in electronic science and technology from Chongqing University, Chongqing, China, in 2012, and the master's degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015.

He is currently a Software Engineer with IBM Research China, Beijing. His research interests include machine learning, recommendation systems, and Bayesian methods.



Yongzhen Huang (M'11) received the B.E. degree in automation control from the Huazhong University of Science and Technology, Wuhan, China, in 2006, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2011.

In July 2011, he joined the National Laboratory of Pattern Recognition, CASIA, where he is currently an Associate Professor. His current research interests include pattern recognition, computer vision, machine learning, and biologically inspired vision computing.

Liang Wang (M'09–SM'09) received the B.E. degree in electronics engineering and the M.E. degree in circuits and systems from Anhui University, Hefei, China, in 1997 and 2000, respectively, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2004.

He is currently a Full Professor of the Hundred Talents Program with the National Laboratory of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, computer vision, and data mining.

Tieniu Tan (F'03) received the B.Sc. degree from Xi'an Jiaotong University, Xi'an, China, in 1984, and the M.Sc. and Ph.D. degrees from Imperial College London, London, U.K., in 1986 and 1989, respectively, all in electronic engineering.

In January 1998, he joined the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he is currently a Professor and the former Director (1998–2013) of the NLPR and Center for Research on Intelligent Perception and Computing. He is currently the Vice President of the Chinese Academy of Sciences. His current research interests include biometrics, image and video understanding, and information content security.