Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Scientific articles recommendation with topic regression and relational matrix factorization*#

Ming YANG, Ying-ming LI, Zhongfei (Mark) ZHANG
(*Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China*)
E-mail: cauchym@zju.edu.cn; liymn@zju.edu.cn; zhongfei@zju.edu.cn

Received Dec. 22, 2013; Revision accepted May 5, 2014; Crosschecked Oct. 15, 2014

Abstract: In this paper we study the problem of recommending scientific articles to users in an online community with a new perspective of considering topic regression modeling and articles relational structure analysis simultaneously. First, we present a novel topic regression model, the topic regression matrix factorization (tr-MF), to solve the problem. The main idea of tr-MF lies in extending the matrix factorization with a probabilistic topic modeling. In particular, tr-MF introduces a regression model to regularize user factors through probabilistic topic modeling under the basic hypothesis that users share similar preferences if they rate similar sets of items. Consequently, tr-MF provides interpretable latent factors for users and items, and makes accurate predictions for community users. To incorporate the relational structure into the framework of tr-MF, we introduce relational matrix factorization. Through combining tr-MF with the relational matrix factorization, we propose the topic regression collective matrix factorization (tr-CMF) model. In addition, we also present the collaborative topic regression model with relational matrix factorization (CTR-RMF) model, which combines the existing collaborative topic regression (CTR) model and relational matrix factorization (RMF). From this point of view, CTR-RMF can be considered as an appropriate baseline for tr-CMF. Further, we demonstrate the efficacy of the proposed models on a large subset of the data from CiteULike, a bibliography sharing service dataset. The proposed models outperform the state-of-the-art matrix factorization models with a significant margin. Specifically, the proposed models are effective in making predictions for users with only few ratings or even no ratings, and support tasks that are specific to a certain field, neither of which has been addressed in the existing literature.

Key words: Matrix factorization, Probabilistic topic modeling, Relational matrix factorization, Recommender system
doi:10.1631/jzus.C1300374 **Document code:** A **CLC number:** TP391

1 Introduction

With the rapid development of social networks, recent years have witnessed fast booming social network sites for academia services, such as CiteULike (http://www.citeulike.org) and Mendeley (http://www.mendeley.com). These social network sites allow researchers to create their personal libraries for the online papers that interest them and to share the libraries with other researchers. This makes recommender systems helpful for researchers to find interesting papers. A recommender system is able to discover and recommend desired scientific articles that researchers have not yet noticed. Thus, it provides a new way for researchers to find interesting articles other than just simple reference tracking or keyword searching.

The traditional approach to recommendation for websites is based on collaborative filtering (CF),

* Project supported by the National Basic Research Program (973) of China (No. 2012CB316400), Zhejiang University–Alibaba Financial Joint Lab, Zhejiang Provincial Engineering Center on Media Data Cloud Processing and Analysis, and the National Science Foundation of the United States (Nos. IIS-0812114 and CCF-1017828)
# A preliminary version was presented at the 22nd ACM International Conference on Information and Knowledge Management, 2013, Burlingame, CA, USA

which is a process of evaluating items through the opinions of other users. Among the existing CF algorithms, matrix factorization models play an important role. Given a rating matrix, the main idea of matrix factorization models is to predict the missing entry (i, j) with an inner product between latent feature vectors for user i (often referred to as user factors) and for item j (often referred to as item factors). While typically the matrix factorization models suffer from the overfitting problem due to the imbalance of the real-world datasets, a feasible solution to this problem is to regularize the latent feature vectors in the matrix factorization models through appropriate priors, e.g., Gaussian prior (Salakhutdinov and Mnih, 2007) and other flexible priors (Salakhutdinov and Mnih, 2008; Ge *et al.*, 2011). Recent work (Agarwal and Chen, 2009; Shan and Banerjee, 2010; Wang and Blei, 2011) improved rating predictions through regularizing item factors with item topic proportions which are captured by the topic modeling. For example, Shan and Banerjee (2010) used the topic proportions in place of the item factors and Wang and Blei (2011) considered the topic proportions as the prior mean of item factors.

However, these models suffer from several limitations. First, they may have met trouble in making accurate predictions for users with only few ratings, which is also a well-known notorious issue for most existing recommender systems. For most of the matrix factorization models, the latent feature vectors for users with only few ratings are close to the prior mean and, consequently, the predicted ratings for the users are seriously influenced by other users. Thus, inappropriate recommendation is provided by such recommender systems, which is typical for scientific article recommender systems. For example, popular articles related to the basic theory of chemistry would be undesirably recommended to computer science researchers who have no interest at all in chemistry, but had few ratings in the history.

This is related to the well-known cold-start problem in the recommender system literature. In general, there are two types of cold-start problems for recommending scientific articles. The first is for new articles that no one has yet rated. The second is for new users whose preferences are almost unknown for a recommender system. The existing literature (Wang and Blei, 2011) for recommending scientific articles focused on the former when they addressed the cold-start problem. We focus on the latter, which makes our work different from the existing literature on this problem.

Second, the methods mentioned above do not effectively support tasks that are specific to a certain research field. Recent work regularizing item factors through latent Dirichlet allocation (LDA) (Blei *et al.*, 2003) could provide a topic representation of articles, but may not effectively support tasks such as 'recommending papers that my papers should cite'. For example, for a biological scientist who is interested but has no expertise at all in data mining, she might wish the recommender systems to support the tasks such as 'recommending popular papers in data mining with potential applications to biological science'.

Third, there is no consideration for the relational structures. The relational structures have become popular in scientific articles, such as citation networks. Furthermore, the relational structures can be approximated by transforming the feature data under a certain distance function, such as the k-nearest neighbor graph. Relational information usually provides useful directions for researchers to find interesting articles. From this point of view, the traditional matrix factorization models that ignore the articles with a relational structure may not be appropriate. On the other hand, recommender systems have been usually applied to cases where user ratings are very sparse. Relational structure information could be considered as the valuable auxiliary information to mitigate this effect, since relational structures provide useful information about the articles.

To address these difficult issues, we first present topic regression matrix factorization (tr-MF), a novel matrix factorization model for the task of recommending scientific articles to users in an online community. The key idea of tr-MF is not only to regularize item factors through a Gaussian prior, but also to regularize user factors through LDA. Given a bag-of-word representation for articles, the idea in LDA is that articles are represented as the topic proportions of the latent topics, where each topic is characterized by a distribution of the words of the articles (here it is emphasized that the article topic proportions are just used for the estimation of the user topic proportions and are not direct contributions to the item factors). Consequently, we may generate user topic proportions by averaging the rated article topic proportions. Treating user topic proportions as the prior mean of the latent feature vectors for users, we assign similar priors to the latent feature vectors for the users who have rated the similar sets of articles. This special way of regularization of user factors has a strong effect on the users with only few ratings as the predictions for these users are more related to the content of the articles. Furthermore, it also makes this method distinct from the recent work (Shan and Banerjee, 2010; Wang and Blei, 2011) that only regularized item factors through LDA.

Furthermore, to effectively incorporate relational structures into the tr-MF algorithm, we introduce relational structure matrix factorization which has been widely applied in the relational data analysis. Consequently, we propose a topic regression collective matrix factorization (tr-CMF) model, which integrates relational structures and tr-MF through topic modeling and collective matrix factorization. We also present the collaborative topic regression model with relational matrix factorization (CTR-RMF), which combines the existing collaborative topic regression (CTR) model (Wang and Blei, 2011) and relational matrix factorization. From this point of view, CTR-RMF can be considered as an appropriate baseline for tr-CMF. Specifically, we connect the relational structure data and user rating data through the shared item factors in the proposed models. The matrix factorization of relational structures helps lead to more accurate item factors for a recommender system.

We conduct the comprehensive evaluations to investigate the performance of the proposed models. The experimental results on the CiteULike dataset demonstrate that the proposed models achieve a significant improvement over the state-of-the-art methods in the performance evaluations. More importantly, our models can provide useful insights into how much the relation information can help improve the prediction performance. Moreover, the tr-MF model is applied to finding the most popular articles for a given topic. The comparison with the state-of-the-art methods demonstrates the promising topic discovery capability of the proposed models.

2 Notations

In this paper we introduce the topic modeling to multiple matrix factorization methods and adopt a bag-of-word representation for items. Suppose there are N users, M items, integer rating values in $\{0, 1\}$, and relation values within the range $[0, 1]$. Let R_{ij} represent the rating of user i for item j: $R_{ij} = 1$ means that user i includes article j in her library; $R_{ij} = 0$ means that article j is not included in user i's library. Let X_{qj} represent the relation of item j for item q: a high value of X_{qj} means that item j has a stronger connection (similarity) with item q than the other items. Let $\boldsymbol{U} \in \mathbb{R}^{K\times N}$ and $\boldsymbol{V} \in \mathbb{R}^{K\times M}$ be the latent feature matrices of users and items, respectively, with column vectors $\boldsymbol{U}_i$ and $\boldsymbol{V}_j$ representing the ith user-specific and jth item-specific latent feature vectors, respectively. Let $\boldsymbol{S} \in \mathbb{R}^{K\times M}$ be the co-item latent feature matrix for relational matrix factorization with column vector $\boldsymbol{S}_q$ representing the qth co-item latent feature vector. Specifically, in the field of scientific article recommendation, every item represents a specific scientific article and its words form a document. A word is denoted as a unit basis vector of size W with exactly one non-zero entry representing the membership to only one word in a dictionary of W words. A document is a collection of P words' occurrences and w_{mp} denotes the pth word of the mth document.

3 Related work

Recommender systems have been extensively studied in the past few years (primarily due to the Netflix Prize). Collaborative filtering is popular and widely used for its domain-free characteristic (Koren *et al.*, 2009). The two primary types of collaborative filtering are the neighborhood based methods (Bell and Koren, 2007; Koren, 2008) and the latent factor models (Bell *et al.*, 2007; Salakhutdinov and Mnih, 2007; 2008; Jiang *et al.*, 2012; Weston *et al.*, 2012). Neighborhood based methods make predictions via computing the relationships between users or between items. Latent factor models try to explain the observed ratings using the latent factors. Among the latent factor models, matrix factorization (MF) models perform well in several real-world applications (Bell *et al.*, 2007; Salakhutdinov and Mnih, 2007; Koren *et al.*, 2009).

In addition, several methods have been proposed to improve the performance of collaborative filtering using the content features (Melville *et al.*, 2002; Yu *et al.*, 2003; Si and Jin, 2004; Jin *et al.*, 2005). In particular, recent work combined the matrix factorization with the topic modeling (Blei *et al.*, 2003; Lafferty and Blei, 2005; Mcauliffe and Blei, 2007) to achieve a better performance (Agarwal and Chen, 2010; Shan and Banerjee, 2010; Wang and Blei, 2011; Purushotham *et al.*, 2012). These models regularize the item factors with the item topic proportions which are captured by the topic modeling. In this work, we regularize the item factors with a Gaussian prior at the same time as regularizing the user factors with the user topic proportions through LDA, which distinguishes all existing models in the methodology.

In the following, we first give a brief review of the matrix factorization models. Then we give a particular description of CTR model (Wang and Blei, 2011), which combines matrix factorization with topic modeling to improve the performance of a recommender system.

3.1 Matrix factorization models

Various matrix factorization (MF) methods have been proposed for collaborative filtering (Srebro *et al.*, 2004; Paterek, 2007; Koren *et al.*, 2009). In the early work, singular value decomposition (SVD) was used for the low rank approximation based on minimizing the sum-squared distance. It finds the matrix $\tilde{\boldsymbol{R}} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}$ of the given rank which minimizes the sum-squared distance to the target matrix $\boldsymbol{R}$. Furthermore, considering that the rating matrix is very sparse in real-world applications, the weighted schemes were proposed in Srebro and Jaakkola (2003). The generalized weighted scheme aims at approximating $\boldsymbol{R}$ with $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}$ and minimizing the objective of a weighted Frobenius loss function as follows:

$$\mathcal{L}(\boldsymbol{U},\boldsymbol{V}) = \sum_{i,j} c_{ij}(R_{ij} - \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2, \tag{1}$$

where c_{ij} is a weighting parameter. In general, it is assigned to 1 for the observed entries and to 0 for the unobserved entries.

To prevent overfitting, a regularization term is appended to the above objective function:

$$\begin{aligned}\mathcal{L}(\boldsymbol{U},\boldsymbol{V}) =& \sum_{i,j} c_{ij}(R_{ij} - \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2 \\ &+ \lambda_U \mathrm{tr}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}) + \lambda_V \mathrm{tr}(\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}),\end{aligned} \tag{2}$$

where λ_U and λ_V are the regularization parameters, and $\mathrm{tr}(\cdot)$ is the trace operator.

This MF model for collaborative filtering is a special case of the regularized low-rank approximation in Salakhutdinov and Mnih (2007). If we denote $\boldsymbol{I}_K$ as the identity matrix with dimension K, then the MF model may be generalized to a probabilistic framework and the generative process is described as follows:

1. For user i, draw $\boldsymbol{U}_i \sim \mathcal{N}(\boldsymbol{0}, \lambda_U^{-1}\boldsymbol{I}_K)$.
2. For item j, draw $\boldsymbol{V}_j \sim \mathcal{N}(\boldsymbol{0}, \lambda_V^{-1}\boldsymbol{I}_K)$.
3. For rating matrix entry (i,j), draw $R_{ij} \sim \mathcal{N}(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, c_{ij}^{-1})$.

When the MF models are applied to tackling the one-class collaborative filtering problems (Hu *et al.*, 2008; Pan *et al.*, 2008), such as page visitation, webpage bookmarking, and news article recommendation, the naive scheme for c_{ij} assigned as '1' to an observed example ($R_{ij} = 1$) and as '0' to an unobserved example ($R_{ij} = 0$) makes mistakes, since the unobserved examples are uncertain. There are two possible explanations for this uncertainty in scientific article recommendations. First, an article is of a user's interest but not read by the user; second, an article is read by a user but not of his or her interest. Typically, an observed example indicates that user i likes article j.

To address this issue, low weights on the error terms were used in Pan *et al.* (2008). In our work, the same strategy setting is used with different non-zero weighting parameters c_{ij} for different ratings R_{ij},

$$c_{ij} = \begin{cases} a, & R_{ij} = 1, \\ b, & R_{ij} = 0, \end{cases} \tag{3}$$

where a and b are the tuning parameters satisfying $a > b > 0$.

3.2 Collaborative topic regression model

Wang and Blei (2011) have proposed the collaborative topic regression (CTR) model to do collaborative filtering based on topic modeling methods. CTR combines the merits of matrix factorization and probabilistic topic modeling methods. CTR represents users with the topic interests and assumes that documents are generated by a topic model. CTR regularizes the item latent factor through the topic proportion. In particular, CTR includes a latent variable ϵ_m that offsets the topic proportions $\boldsymbol{\theta}_m$

when modeling the user ratings. This offset variable can be considered as a kind of social network impact.

The CTR model has achieved a good performance in using content information for item recommendation. However, it has trouble in making accurate predictions for users who have only few ratings, since latent user factors are close to the prior mean when users' ratings are few. Furthermore, there is no consideration for article relational structures in CTR. It has been well studied and established in the document analysis research area that article relational structures provide directions for users to find interesting articles. In our work, we propose novel models to address the above existing problems.

4 Topic regression matrix factorization for recommending scientific articles

4.1 Topic regression matrix factorization model

Under the assumption that users have similar preferences if they have rated similar sets of articles, we introduce a regression model to regularize user factors through LDA.

Our method is based on placing the constraint prior mean on the latent feature vectors for users. In particular, we first obtain the set of item topic proportions $\boldsymbol{\theta}=(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_M)$ through LDA. In LDA, each $\boldsymbol{\theta}_m$ follows the Dirichlet distribution for parameter α ($\mathrm{Dir}(\alpha)$); each z_{mp} is the topic for the pth word in mth document follows the multinomial distribution with parameter $\boldsymbol{\theta}_m$ ($\mathrm{Mult}(\boldsymbol{\theta}_m)$); each word w_{mp} follows the multinomial distribution with parameter $\boldsymbol{\beta}_{z_{mp}}$ ($\mathrm{Mult}(\boldsymbol{\beta}_{z_{mp}})$) where $\boldsymbol{\beta}$ is a matrix, each row of which denotes the word distribution of a topic. Consequently, we generate user topic proportions for user i:

$$\tilde{\boldsymbol{\theta}}_i = \frac{\sum_{m=1}^{M} J_{im}\boldsymbol{\theta}_m}{\sum_{m=1}^{M} J_{im}}, \tag{4}$$

where $\boldsymbol{J}$ is the observed indicator matrix with J_{im} taking on value 1 if user i has rated item m and 0 otherwise. Then, we define the latent feature vector for user i as

$$\boldsymbol{U}_i = \boldsymbol{Y}_i + \tilde{\boldsymbol{\theta}}_i. \tag{5}$$

Here $\boldsymbol{Y}_i$ is considered as the offset added to $\tilde{\boldsymbol{\theta}}_i$ to obtain the latent feature vector $\boldsymbol{U}_i$. Specifically, when $\boldsymbol{Y}_i \sim \mathcal{N}(\mathbf{0}, \lambda_U^{-1}\boldsymbol{I}_K)$, we obtain the distribution of $\boldsymbol{U}_i$:

$$\boldsymbol{U}_i \sim \mathcal{N}(\tilde{\boldsymbol{\theta}}_i, \lambda_U^{-1}\boldsymbol{I}_K). \tag{6}$$

Thus, $\tilde{\boldsymbol{\theta}}_i$ now is considered as the prior mean of the latent feature vector $\boldsymbol{U}_i$. Specifically, in the probabilistic matrix factorization (PMF) model (Salakhutdinov and Mnih, 2007), $\boldsymbol{U}_i$ and $\boldsymbol{Y}_i$ are equal because the prior mean is fixed to zero.

In the process of filtering, we place a zero-mean Gaussian prior to the latent feature vector for item j:

$$\boldsymbol{V}_j \sim \mathcal{N}(\mathbf{0}, \lambda_V^{-1}\boldsymbol{I}_K). \tag{7}$$

We now define the rating that user i gives item j as follows:

$$R_{ij} \sim \mathcal{N}(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, c_{ij}^{-1}). \tag{8}$$

For an easy reference, the graphical model of tr-MF is shown in Fig. 1. Furthermore, the particular generative process of tr-MF is described as follows:

1. **For** document $m = 1, 2, \ldots, M$
   - Draw $\boldsymbol{\theta}_m \sim \mathrm{Dir}(\alpha)$;
   - **For** word $p = 1, 2, \ldots, P_m$
     - Draw $z_{mp} \sim \mathrm{Mult}(\boldsymbol{\theta}_m)$;
     - Draw $w_{mp} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{mp}})$;
2. **For** item $j = 1, 2, \ldots, M$
   - Draw $\boldsymbol{V}_j \sim \mathcal{N}(\mathbf{0}, \lambda_V^{-1}\boldsymbol{I}_K)$;
3. **For** user $i = 1, 2, \ldots, N$
   - Draw $\boldsymbol{Y}_i \sim \mathcal{N}(\mathbf{0}, \lambda_U^{-1}\boldsymbol{I}_K)$;
   - Compute $\boldsymbol{U}_i = \boldsymbol{Y}_i + \tilde{\boldsymbol{\theta}}_i$;
4. **For** user $i = 1, 2, \ldots, N$
   - **For** item $j = 1, 2, \ldots, M$
     - Draw $R_{ij} \sim \mathcal{N}(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, c_{ij}^{-1})$.

4.2 Discussion

The key property in tr-MF lies in how the user factor is generated. It is shown in Eq. (5) that the user factor $\boldsymbol{U}_i$ is a mixture of two sources: (1) the user topic proportions $\tilde{\boldsymbol{\theta}}_i$ which is obtained by averaging the rated item topic proportions; (2) the offset vector $\boldsymbol{Y}_i$ which reflects the influence from other users' ratings. This perspective actually accounts for the process of making a prediction: the recommender system first learns the knowledge from the user's rated items and then combines the suggestions from other users with what it learned through the collaborative filtering to form the accurate prediction. Therefore, the predictions count more on

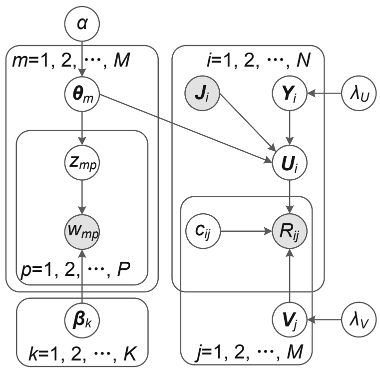

Fig. 1 Graphical model for tr-MF. In the theory of the graphical model, each circle node in the graph corresponds to a random variable and the shaded one denotes an observed variable. Rectangles denote replication of the model within the rectangle. The indexes and the ranges of the replicates are given inside the rectangle at bottom or top

the content of their rated items. Herein, our model recommends the articles concentrating on the area of interest of the researchers.

4.3 Parameter estimation

We aim at maximizing a posterior of $\boldsymbol{U}$, $\boldsymbol{V}$, and $\boldsymbol{\theta}$, given the training data for a precise prediction, which is equivalent to maximizing the log likelihood of $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{\theta}$, and $\boldsymbol{R}$ given the hyper-parameters λ_U, λ_V, and $\boldsymbol{\beta}$. We adopt the expectation maximization (EM) algorithm to learn the parameters of tr-MF. The likelihood is denoted as

$$\begin{aligned}\mathcal{L} = &-\sum_{i,j=1}^{N,M} c_{ij}(R_{ij} - \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2 \\ &-\lambda_U \sum_{i=1}^{N} ||\boldsymbol{U}_i - \tilde{\boldsymbol{\theta}}_i||^2 - \lambda_V \sum_{j=1}^{M} ||\boldsymbol{V}_j||^2 \\ &+ 2\sum_m \sum_p \log\left(\sum_k \theta_{mk}\beta_{k,z_{mp}}\right) + \text{constant} \\ = &-\|\sqrt{\boldsymbol{C}} \odot (\boldsymbol{R} - \boldsymbol{U}^{\mathrm{T}}\boldsymbol{V})\|_F^2 - \lambda_U\|\boldsymbol{U} - \tilde{\boldsymbol{\theta}}\|_F^2 \\ &-\lambda_V\|\boldsymbol{V}\|_F^2 + 2\sum_m \sum_p \log\left(\sum_k \theta_{mk}\beta_{k,z_{mp}}\right) \\ &+ \text{constant}, \end{aligned} \tag{9}$$

where $\odot$ indicates the Hadamard product of two matrices, $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2, \ldots, \tilde{\boldsymbol{\theta}}_N)$, and $\sqrt{\boldsymbol{C}}$ indicates the element-wise square root of the matrix $\boldsymbol{C}$. For simplicity, the hyper-parameter of the Dirichlet distribution α is assigned to 1 in tr-MF. Notice that $\boldsymbol{C} \odot \boldsymbol{R} = \boldsymbol{R}$. Setting the derivatives with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ to zero, we obtain the updating equations for $\boldsymbol{U}$ and $\boldsymbol{V}$ in a vectorization form:

$$\begin{aligned}\mathrm{vec}(\boldsymbol{U}) &= (\lambda_U \boldsymbol{I}_{KN} + \tilde{\boldsymbol{V}}\boldsymbol{D}_1\tilde{\boldsymbol{V}}^{\mathrm{T}})^{-1}\mathrm{vec}(\boldsymbol{V}\boldsymbol{R}^{\mathrm{T}} + \lambda_U\tilde{\boldsymbol{\theta}}), \\ \mathrm{vec}(\boldsymbol{V}) &= (\lambda_V \boldsymbol{I}_{KM} + \tilde{\boldsymbol{U}}\boldsymbol{D}_2\tilde{\boldsymbol{U}}^{\mathrm{T}})^{-1}\mathrm{vec}(\boldsymbol{U}\boldsymbol{R}),\end{aligned}$$

where $\tilde{\boldsymbol{V}} = \boldsymbol{I}_N \otimes \boldsymbol{V}$, $\tilde{\boldsymbol{U}} = \boldsymbol{I}_M \otimes \boldsymbol{U}$, $\boldsymbol{D}_1 = \mathrm{diag}\left(\mathrm{vec}(\boldsymbol{C}^{\mathrm{T}})\right)$, and $\boldsymbol{D}_2 = \mathrm{diag}\left(\mathrm{vec}(\boldsymbol{C})\right)$. $\otimes$ is the Kronecker product of two matrices, $\mathrm{vec}(\cdot)$ is the vectorization function mapping a matrix to a column vector stacked by the columns of the matrix, for example $\mathrm{vec}(\boldsymbol{U}) = (\boldsymbol{U}_1^{\mathrm{T}}, \boldsymbol{U}_2^{\mathrm{T}}, \ldots, \boldsymbol{U}_N^{\mathrm{T}},)^{\mathrm{T}}$, and $\mathrm{diag}(\cdot)$ is the diagonal function mapping a vector to a diagonal matrix. Since there is no closed-form solution in Eq. (9) for updating $\boldsymbol{\theta}$, we implement an EM framework to learn $\boldsymbol{\theta}$. We define $\mathcal{L}_\theta$ as the component in Eq. (9) associated with $\boldsymbol{\theta}$:

$$\mathcal{L}_\theta = -\lambda_U\|\boldsymbol{U} - \tilde{\boldsymbol{\theta}}\|_F^2 + 2\sum_m \sum_p \log\left(\sum_k \theta_{mk}\beta_{k,z_{mp}}\right),$$

and we induce an auxiliary distribution q and a latent variable $\phi_{mpk} = q(z_{mp} = k)$ to formulate a lower bound $Q(\boldsymbol{\theta}, \boldsymbol{\phi})$ of $\mathcal{L}_\theta$:

$$\begin{aligned}Q(\boldsymbol{\theta}, \boldsymbol{\phi}) = &-\lambda_U\|\boldsymbol{U} - \tilde{\boldsymbol{\theta}}\|_F^2 \\ &+ 2\sum_m \sum_p \sum_k \phi_{mpk} \log\left(\frac{\theta_{mk}\beta_{k,z_{mp}}}{\phi_{mpk}}\right).\end{aligned} \tag{10}$$

Therefore, the EM framework is implemented to make the lower bound $Q(\boldsymbol{\theta}, \boldsymbol{\phi})$ approximate to $\mathcal{L}_\theta$.

E-step: From the Bayes' formula, we obtain a posterior probability of z_{mp}:

$$\phi_{mpk} = \frac{\theta_{mk}\beta_{k,z_{mp}}}{\sum_k \theta_{mk}\beta_{k,z_{mp}}}.$$

M-step: Maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\phi})$ with the constraints that $\sum_k \theta_{mk} = 1$ and $\sum_p \beta_{k,z_{mp}} = 1$, we obtain

$$\beta_{k,w} = \frac{\sum_{m,p} \phi_{mpk} 1[w_{mp} = w]}{\sum_{m,p,k} \phi_{mpk} 1[w_{mp} = w]}, \tag{11}$$

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \theta_{mk}} = &2\lambda_U \sum_{i=1}^{N} \frac{(U_{ik} - \tilde{\theta}_{mk})J_{im}}{\sum_{m=1}^{M} J_{im}} \\ &+ \frac{2\sum_p \phi_{mpk}}{\theta_{mk}} + 2\lambda,\end{aligned} \tag{12}$$

where λ is the Lagrangian multiplier and note that $\boldsymbol{\theta}$ cannot be updated in a closed form; hence, we use the projection gradient to learn $\boldsymbol{\theta}$. The term 1[condition] is an indicator function. When the condition is true, the function value is 1 and otherwise it is 0.

4.4 Making prediction

We apply the updated tr-MF model to computing the likelihood and updating $\boldsymbol{U}$, $\boldsymbol{V}$, $\tilde{\boldsymbol{\theta}}$, and $\boldsymbol{\beta}$, which can be used to make predictions after the iteration converges and the tr-MF model is established. We predict R_{ij} from its expectation:

$$E[R_{ij}|\boldsymbol{Y}_i, \boldsymbol{V}_j, \boldsymbol{\theta}, \boldsymbol{J}_i] = {\boldsymbol{U}_i}^{\mathrm{T}}\boldsymbol{V}_j = (\boldsymbol{Y}_i + \tilde{\boldsymbol{\theta}}_i)^{\mathrm{T}}\boldsymbol{V}_j. \quad (13)$$

For new users with no ratings in the cold-start scenarios, $E[\boldsymbol{Y}_i] = \boldsymbol{0}$ and we predict the rating with

$$E[R_{ij}|\boldsymbol{V}_j, \boldsymbol{\theta}, \boldsymbol{J}_i] = \tilde{\boldsymbol{\theta}}_i^{\mathrm{T}}\boldsymbol{V}_j. \quad (14)$$

We obtain the user topics $\tilde{\boldsymbol{\theta}}_i$ through the topic analysis of their activities, such as browsing histories, search histories, or comments.

5 Collaborative filtering with relational structure information

Since the relational structure information is the valuable auxiliary information for a recommender system, we take advantage of this information to improve the generalization performance of matrix factorization models. In this section, we first propose a relational matrix factorization model. Then we update the framework of CTR and tr-MF including this model, respectively.

There are many ways to combine the relational matrix factorization with tr-MF or CTR for scientific articles recommendation. Our idea is to fuse them into a single, consistent, and compact representation. To achieve this goal, we connect the relational structure data and user ratings data through shared item factors. Specifically, we propose two novel models called collaborative topic regression with relational matrix factorization (CTR-RMF) and topic regression collective matrix factorization (tr-CMF), respectively. Based on the proposed models, we may see how much articles relational information could help improve the prediction performance.

5.1 Probabilistic relational matrix factorization

To take advantage of the item (article) relational structures, we propose a probabilistic relational matrix factorization model which aims at approximating item relational matrix $\boldsymbol{X}$ with $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}$.

In the process of factorization, we place a zero-mean Gaussian prior to the co-item latent feature vector for item q:

$$\boldsymbol{S}_q \sim \mathcal{N}(\boldsymbol{0}, \lambda_S^{-1}\boldsymbol{I}_K). \quad (15)$$

The latent feature vector for item j is similar to that in the last section:

$$\boldsymbol{V}_j \sim \mathcal{N}(\boldsymbol{0}, \lambda_V^{-1}\boldsymbol{I}_K). \quad (16)$$

We now define the relation value between items q and j as follows:

$$X_{qj} \sim \mathcal{N}(\boldsymbol{S}_q^{\mathrm{T}}\boldsymbol{V}_j, \lambda_X^{-1}). \quad (17)$$

The logarithm of the posterior distribution over $\boldsymbol{S}$ and $\boldsymbol{V}$ is given by

$$\begin{aligned}&\ln p(\boldsymbol{S}, \boldsymbol{V}|\boldsymbol{X}, \lambda_X, \lambda_S, \lambda_V)\\ &= -\sum_{q,j=1}^{M,M}\lambda_X(X_{qj} - \boldsymbol{S}_q^{\mathrm{T}}\boldsymbol{V}_j)^2 - \lambda_S\sum_{q=1}^{M}||\boldsymbol{S}_q||^2\\ &\quad -\lambda_V\sum_{j=1}^{M}||\boldsymbol{V}_j||^2 + \text{constant}, \quad (18)\end{aligned}$$

where λ_X and λ_S are the regularization parameters.

5.2 Collaborative topic regression model with relational matrix factorization

The main idea of CTR-RMF is as follows. For each item m, $\boldsymbol{V}_m$ not only serves as RMF's item latent feature vector for its corresponding relationship, but also serves as CTR's item latent feature vector. Therefore, the common $\boldsymbol{V}_m$ for both ratings and relational information of article m becomes the glue to combine RMF and CTR. In particular, the graphical model of CTR-RMF is shown in Fig. 2. Furthermore, the particular generative process of CTR-RMF is described as follows:

1. **For** document $m = 1, 2, \ldots, M$
   Draw $\boldsymbol{\theta}_m \sim \mathrm{Dir}(\alpha)$;
   For word $p = 1, 2, \ldots, P_m$
      Draw $z_{mp} \sim \mathrm{Mult}(\boldsymbol{\theta}_m)$;
      Draw $w_{mp} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{mp}})$;
2. **For** item $m = 1, 2, \ldots, M$
   Draw $\boldsymbol{V}_m \sim \mathcal{N}(\boldsymbol{0}, \lambda_V^{-1}\boldsymbol{I}_K)$;
3. **For** user $i = 1, 2, \ldots, N$
   Draw $\boldsymbol{U}_i \sim \mathcal{N}(\boldsymbol{0}, \lambda_U^{-1}\boldsymbol{I}_K)$;
4. **For** user $i = 1, 2, \ldots, N$
   For item $m = 1, 2, \ldots, M$

Draw $R_{im} \sim \mathcal{N}(\boldsymbol{U}_i^{\mathrm{T}} \boldsymbol{V}_m, c_{im}^{-1})$;

5. **For** co-item $q = 1, 2, \ldots, M$

Draw $\boldsymbol{S}_q \sim \mathcal{N}(\boldsymbol{0}, \lambda_S^{-1} \boldsymbol{I}_K)$;

6. **For** co-item $q = 1, 2, \ldots, M$

for item $m = 1, 2, \ldots, M$

Draw $X_{qm} \sim \mathcal{N}(\boldsymbol{S}_q^{\mathrm{T}} \boldsymbol{V}_m, \lambda_X^{-1})$.

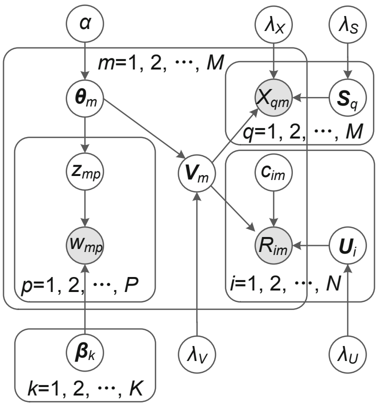

Fig. 2 Graphical model for CTR-RMF

We adopt the EM algorithm to learn the parameters of CTR-RMF. The likelihood is denoted as

$$\begin{aligned}
\mathcal{L} =& - \sum_{i,m=1}^{N,M} c_{ij} (R_{im} - \boldsymbol{U}_i^{\mathrm{T}} \boldsymbol{V}_m)^2 \\
& - \lambda_U \sum_{i=1}^{N} ||\boldsymbol{U}_i||^2 - \lambda_V \sum_{m=1}^{M} ||\boldsymbol{V}_m - \boldsymbol{\theta}_m||^2 \\
& - \lambda_X \sum_{q,m=1}^{M,M} (X_{qm} - \boldsymbol{S}_q^{\mathrm{T}} \boldsymbol{V}_m)^2 - \lambda_S \sum_{q=1}^{M} ||\boldsymbol{S}_q||^2 \\
& + 2 \sum_m \sum_p \log \left(\sum_k \theta_{mk} \beta_{k,z_{mp}} \right) + \text{constant} \\
=& - \| \sqrt{\boldsymbol{C}} \odot (\boldsymbol{R} - \boldsymbol{U}^{\mathrm{T}} \boldsymbol{V}) \|_F^2 \\
& - \lambda_U \|\boldsymbol{U}\|_F^2 - \lambda_V \|\boldsymbol{V} - \boldsymbol{\theta}\|_F^2 \\
& - \lambda_X \|\boldsymbol{X} - \boldsymbol{S}^{\mathrm{T}} \boldsymbol{V}\|_F^2 - \lambda_S \|\boldsymbol{S}\|_F^2 \\
& + 2 \sum_m \sum_p \log \left(\sum_k \theta_{mk} \beta_{k,z_{mp}} \right) + \text{constant}.
\end{aligned} \tag{19}$$

For simplicity, α is assigned to 1 in CTR-RMF. Setting the derivatives with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ to zero, we obtain the updating equations for $\boldsymbol{U}$ and $\boldsymbol{V}$ in a vectorization form:

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{U}) =& (\lambda_U \boldsymbol{I}_{KN} + \tilde{\boldsymbol{V}} \boldsymbol{D}_1 \tilde{\boldsymbol{V}}^{\mathrm{T}})^{-1} \mathrm{vec}(\boldsymbol{V} \boldsymbol{R}^{\mathrm{T}}), \\
\mathrm{vec}(\boldsymbol{V}) =& (\lambda_V \boldsymbol{I}_{KM} + \tilde{\boldsymbol{U}} \boldsymbol{D}_2 \tilde{\boldsymbol{U}}^{\mathrm{T}} + \lambda_X \tilde{\boldsymbol{S}} \tilde{\boldsymbol{S}}^{\mathrm{T}})^{-1} \\
& \cdot \mathrm{vec}(\boldsymbol{U}\boldsymbol{R} + \lambda_X \boldsymbol{S}\boldsymbol{X} + \lambda_V \boldsymbol{\theta}), \\
\boldsymbol{S} =& (\lambda_S \boldsymbol{I}_K + \lambda_X \boldsymbol{V}\boldsymbol{V}^{\mathrm{T}})^{-1} (\lambda_X \boldsymbol{V} \boldsymbol{X}^{\mathrm{T}}),
\end{aligned}$$

where $\tilde{\boldsymbol{S}} = \boldsymbol{I}_M \otimes \boldsymbol{S}$. Since there is no closed-form solution in Eq. (19) for updating $\boldsymbol{\theta}$,

$$\mathcal{L}_\theta = -\lambda_V \|\boldsymbol{V} - \boldsymbol{\theta}\|_F^2 + 2 \sum_m \sum_p \log \left(\sum_k \theta_{mk} \beta_{k,z_{mp}} \right),$$

and we induce an auxiliary distribution q and a latent variable $\phi_{mpk} = q(z_{mp} = k)$ to formulate a lower bound $Q(\boldsymbol{\theta}, \boldsymbol{\phi})$ of $\mathcal{L}_\theta$:

$$\begin{aligned}
Q(\boldsymbol{\theta}, \boldsymbol{\phi}) =& - \lambda_V \|\boldsymbol{V} - \boldsymbol{\theta}\|_F^2 \\
& + 2 \sum_m \sum_p \sum_k \phi_{mpk} \log \left(\frac{\theta_{mk} \beta_{k,z_{mp}}}{\phi_{mpk}} \right).
\end{aligned}$$

Therefore, the EM framework is implemented to make the lower bound $Q(\boldsymbol{\theta}, \boldsymbol{\phi})$ approximate to $\mathcal{L}_\theta$.

E-step: From the Bayes' formula, we obtain a posterior probability of z_{mp}:

$$\phi_{mpk} = \frac{\theta_{mk} \beta_{k,z_{mp}}}{\sum_k \theta_{mk} \beta_{k,z_{mp}}}.$$

M-step: Maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\phi})$ with the constraints that $\sum_k \theta_{mk} = 1$ and $\sum_p \beta_{k,z_{mp}} = 1$, we obtain

$$\beta_{k,w} = \frac{\sum_{m,p} \phi_{mpk} 1[w_{mp} = w]}{\sum_{m,p,k} \phi_{mpk} 1[w_{mp} = w]}, \tag{20}$$

$$\frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\phi})}{\partial \theta_{mk}} = 2\lambda_V (V_{mk} - \theta_{mk}) + \frac{2 \sum_p \phi_{mpk}}{\theta_{mk}} + 2\lambda. \tag{21}$$

Note that $\boldsymbol{\theta}$ cannot be updated in a closed form; hence, we use the projection gradient to learn $\boldsymbol{\theta}$.

5.3 Topic regression collective matrix factorization

The main idea of tr-CMF is as follows. For each item j, $\boldsymbol{V}_m$ not only serves as RMF's item latent feature vector for its corresponding relationship, but also serves as tr-MF's item latent feature vector. Therefore, the common $\boldsymbol{V}_j$ for both ratings and relational information of article j becomes the glue to combine RMF and tr-MF together in a similar way to CTR-RMF.

In particular, the graphical model of tr-CMF is shown in Fig. 3. Furthermore, the particular generative process of tr-CMF is described as follows:

1. **For** document $m = 1, 2, \ldots, M$

    Draw $\boldsymbol{\theta}_m \sim \mathrm{Dir}(\alpha)$;

    For word $p = 1, 2, \ldots, P_m$

        Draw $z_{mp} \sim \mathrm{Mult}(\boldsymbol{\theta}_m)$;

        Draw $w_{mp} \sim \mathrm{Mult}(\boldsymbol{\beta}_{z_{mp}})$;

2. **For** item $j = 1, 2, \ldots, M$

    Draw $\boldsymbol{V}_j \sim \mathcal{N}(\boldsymbol{0}, \lambda_V^{-1}\boldsymbol{I}_K)$;

3. **For** co-item $q = 1, 2, \ldots, M$

    Draw $\boldsymbol{S}_q \sim \mathcal{N}(\boldsymbol{0}, \lambda_S^{-1}\boldsymbol{I}_K)$;

4. **For** user $i = 1, 2, \ldots, N$

    Draw $\boldsymbol{Y}_i \sim \mathcal{N}(\boldsymbol{0}, \lambda_U^{-1}\boldsymbol{I}_K)$; $\boldsymbol{U}_i = \boldsymbol{Y}_i + \tilde{\boldsymbol{\theta}}_i$;

5. **For** user $i = 1, 2, \ldots, N$

    For item $m = 1, 2, \ldots, M$

        Draw $R_{ij} \sim \mathcal{N}(\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j, c_{ij}^{-1})$;

6. **For** co-item $q = 1, 2, \ldots, M$

    Draw $\boldsymbol{S}_q \sim \mathcal{N}(\boldsymbol{0}, \lambda_S^{-1}\boldsymbol{I}_K)$;

7. **For** co-item $q = 1, 2, \ldots, M$

    For item $j = 1, 2, \ldots, M$

        Draw $X_{qj} \sim \mathcal{N}(\boldsymbol{S}_q^{\mathrm{T}}\boldsymbol{V}_j, \lambda_X^{-1})$.

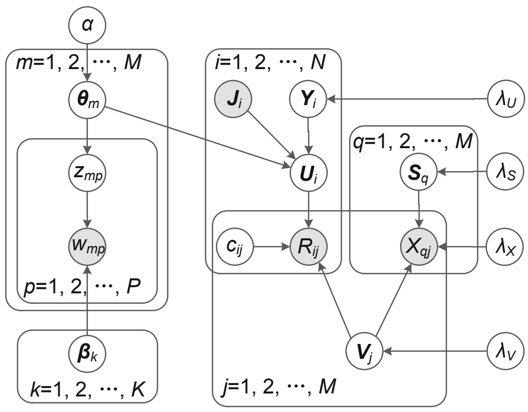

Fig. 3 Graphical model for tr-CMF

We adopt the EM algorithm to learn the parameters of tr-CMF. The likelihood is denoted as

$$\begin{aligned}
\mathcal{L} = &- \sum_{i,j=1}^{N,M} c_{ij}(R_{ij} - \boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j)^2 \\
&- \lambda_U \sum_{i=1}^{N} ||\boldsymbol{U}_i - \tilde{\boldsymbol{\theta}}_i||^2 - \lambda_V \sum_{j=1}^{M} ||\boldsymbol{V}_j||^2 \\
&- \lambda_X \sum_{q,j=1}^{M,M} (X_{qj} - \boldsymbol{S}_q^{\mathrm{T}}\boldsymbol{V}_j)^2 - \lambda_S \sum_{q=1}^{M} ||\boldsymbol{S}_q||^2 \\
&+ 2\sum_m \sum_p \log\left(\sum_k \theta_{mk}\beta_{k,z_{mp}}\right) + \text{constant} \\
= &- \|\sqrt{\boldsymbol{C}} \odot (\boldsymbol{R} - \boldsymbol{U}^{\mathrm{T}}\boldsymbol{V})\|_F^2 \\
&- \lambda_U \|\boldsymbol{U} - \tilde{\boldsymbol{\theta}}\|_F^2 - \lambda_V \|\boldsymbol{V}\|_F^2 \\
&- \lambda_X \|\boldsymbol{X} - \boldsymbol{S}^{\mathrm{T}}\boldsymbol{V}\|_F^2 - \lambda_S \sum_{q=1}^{M} ||\boldsymbol{S}_q||^2 \\
&+ 2\sum_m \sum_p \log\left(\sum_k \theta_{mk}\beta_{k,z_{mp}}\right) + \text{constant}.
\end{aligned} \tag{22}$$

For simplicity, α is assigned to 1 in tr-CMF. Setting the derivatives with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ to zero, we obtain the updating equations for $\boldsymbol{U}$ and $\boldsymbol{V}$ in a vectorization form:

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{U}) &= (\lambda_U \boldsymbol{I}_{KN} + \tilde{\boldsymbol{V}}\boldsymbol{D}_1\tilde{\boldsymbol{V}}^{\mathrm{T}})^{-1}\mathrm{vec}(\boldsymbol{V}\boldsymbol{R}^{\mathrm{T}} + \lambda_U\tilde{\boldsymbol{\theta}}), \\
\mathrm{vec}(\boldsymbol{V}) &= (\lambda_V \boldsymbol{I}_{KM} + \lambda_X\tilde{\boldsymbol{S}}\tilde{\boldsymbol{S}}^{\mathrm{T}} + \tilde{\boldsymbol{U}}\boldsymbol{D}_2\tilde{\boldsymbol{U}}^{\mathrm{T}})^{-1} \\
&\quad \cdot \mathrm{vec}(\boldsymbol{U}\boldsymbol{R} + \lambda_X\boldsymbol{S}\boldsymbol{X}), \\
\boldsymbol{S} &= (\lambda_S \boldsymbol{I}_K + \lambda_X\boldsymbol{V}\boldsymbol{V}^{\mathrm{T}})^{-1}(\lambda_X\boldsymbol{V}\boldsymbol{X}^{\mathrm{T}}).
\end{aligned}$$

For the updating of $\boldsymbol{\theta}$, we adopt a similar strategy to that of tr-MF in Section 4.

6 Experiments

We show the effectiveness of the proposed models using a real-world community dataset.

6.1 Validation dataset

We use a dataset from CiteULike (http://www.citeulike.org/faq/data.adp). CiteULike is a social network website for scientific researchers, and it allows users to create personal reference libraries for the interesting articles and to capture all the metadata (authors, abstract, keywords, etc.) of the articles. In the prior work, Wang and Blei (2011) collected a large subset from CiteULike to form their dataset. For a fair comparison study, we use the same dataset as the benchmark.

This dataset contains 204 986 pairs of observed ratings with 5551 users and 16 980 articles. However, the sparseness is still quite low, i.e., 0.2175%, which is much lower than that of the well-known Movielens dataset (http://www.cs.umn.edu/Research/GroupLens) with the sparseness 4.25%. On average, each user has 37 articles in the library, ranging from 10 to 403, and each article appears in 12 users' libraries, ranging from 1 to 321. For each article, its title and abstract can be used to obtain the bag-of-word representation. The corpus has 8000 distinct words after the standard text processing by term frequency inverse document frequency (tf-idf) and removing the stop words. These articles were added to CiteULike between 2004 and 2010.

For article relational structures, the direct choice is citation information. However, such information is not available for CiteULike. We use the relational structures based on the articles' similarities. In particular, we construct relational data for each item (document) with cosine similarities between the corresponding term-frequency vectors.

6.2 Evaluation metric

Two metrics, the accuracy and recall, are often used to measure the performance of a recommender system. However, as we discussed earlier, in one-class collaborative filtering, especially for recommending scientific articles, the ratings of $R_{ij} = 0$ are uncertain and may have two different explanations: the user does not like an article or does not know it. But the ratings of $R_{ij} = 1$ are known to be true positive. Thus, here we use the recall which only focuses on the true positive examples. Given the number of the recommended items T, the recall@T is defined as

$$\text{recall@}T = \frac{\text{the number of articles the user likes in top } T}{\text{total number of articles the user likes}}.$$

The recall above is user-oriented. To obtain the recall for the entire system, we first compute the recall for each user; and then obtain the overall recall by averaging the recalls from all the users.

6.3 Experimental settings

We use cross-validation to estimate the performance of different algorithms. The validation dataset is randomly divided into training and test sets with a 80/20 splitting ratio. The training set contains 80% known positive examples ($R_{ij} = 1$) and 80% unknown examples ($R_{ij} = 0$). The test set includes the other 20% known positive examples and unknown examples. We form predictive ratings for the test set and generate a list of the top T recommended articles.

The parameter settings for MF (Pan *et al.*, 2008), CTR (Wang and Blei, 2011), tr-MF, CTR-RMF, and tr-CMF models are listed in Table 1. Note that the other parameters for CTR-RMF and tr-CMF models are varied so as to study their effect on prediction results. All the models are compared under the same test set.

6.4 Performance comparisons

Several performance comparisons of all the models are shown in Fig. 4 with the varying number of the recommended articles $T = \{50, 100, \ldots, 300\}$. As is shown, tr-CMF outperforms the competing models with significant margins in almost all the cases, since the recall achieved by tr-CMF is higher than that achieved by the competing models in most cases. First, the recall of tr-CMF improves 17% on average with the increase of T from 50 to 300; this can be explained as that the content restriction for user factors should be more effective when T becomes larger. Second, the performance of tr-CMF is better than those of the competing models with the increase in the latent feature vectors' dimensionality K. Taking $T = 50$ for example, the recall of tr-CMF rises from 0.68 to 0.71, while the recall of MF does not increase significantly as K increases; the recall of CTR is small though by an increase of 5% as K increases; this can be explained by showing that the knowledge of the articles is fully discovered through LDA

Table 1 Parameter settings for different models

Model	λ_U	λ_V	a	b
MF	0.01	0.01	1	0.01
CTR	0.01	100	1	0.01
tr-MF	100	0.01	1	0.01
CTR-RMF	0.01	100	1	0.01
tr-CMF	100	0.01	1	0.01

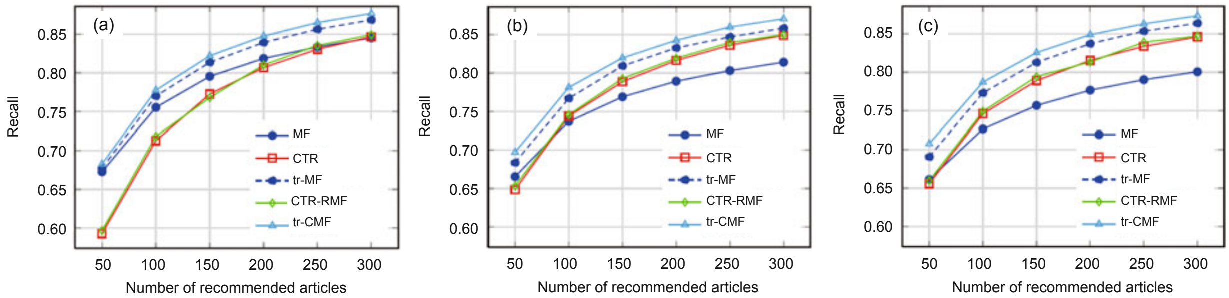

Fig. 4 Comparison of the recall on predictions for different dimensionalities of the latent feature vectors K: (a) $K = 100$; (b) $K = 200$; (c) $K = 300$

when K becomes larger. The above result suggests that it is essential to introduce the topic modeling to enhance the performance of the matrix factorization model.

The experiments also reveal a number of interesting observations.

Both tr-CMF and tr-MF obtain a better performance than the other competing models. This shows the importance of regularizing the user factors with topic modeling.

The fact that tr-CMF significantly outperforms CTR-RMF proves that tr-MF behaves better than CTR in the efficiency of exploiting the article relational structure.

6.5 Sensitivity study of parameters λ_U, λ_X

Fig. 5 shows the comparison study when λ_U varies for tr-MF and tr-CMF while the regularization parameter λ_X for relational structure is fixed. We take CTR as our baseline. As is shown, tr-CMF and tr-MF outperform CTR with significant margins in almost all the cases. When λ_U is around 10, the penalty of user-specific latent feature vector $\boldsymbol{U}_i$ diverging from the user topic proportions $\tilde{\boldsymbol{\theta}}_i$ is less than 1%. When λ_U increases to the region $[10, 100]$, $\tilde{\boldsymbol{\theta}}_i$ plays a more important role in recommending articles and the performance of tr-MF and tr-CMF outperformed CTR with 3% and 5%, respectively. When λ_U is larger than 1000, however, $\boldsymbol{U}_i$ is almost equal to $\tilde{\boldsymbol{\theta}}_i$ and the performance substantially drops. This effect indicates that an appropriate penalty helps improve the prediction performance.

Fig. 6 shows the comparison study when λ_X varies for tr-CMF and CTR-RMF while the user factor regularization parameter λ_U is fixed. As is shown, the values of λ_X have a significant impact on the recommendation results of tr-CMF, which demonstrates that exploiting the relational structure information with tr-MF improves recommendation recall. For values of λ_X less than 0.01, the improvement in recall is small and it increases with the increase in λ_X. When λ_X is increased to 0.1, the performance of tr-CMF increases 1%; when λ_X increases to 1, however, tr-CMF gives greater preference to relational structure information but less preference to

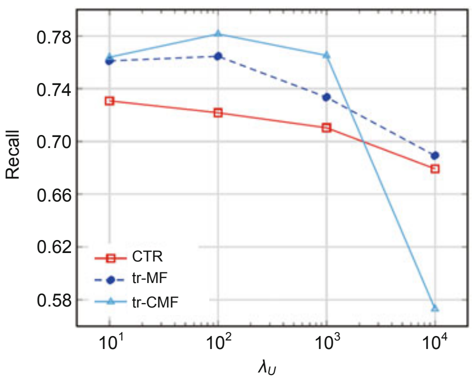

Fig. 5 Comparison of the recalls under different λ_U ($K = 200$, $T = 100$)

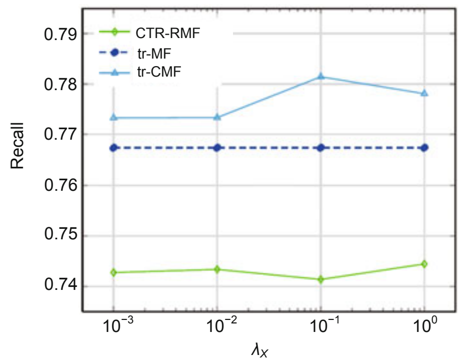

Fig. 6 Comparison of the recalls under different λ_X ($K = 200$, $T = 100$)

user's ratings data and the performance substantially drops. This effect indicates that an appropriate regularization of relational structure information helps improve the prediction performance.

6.6 Performance on different users

For the matrix factorization models, the prediction performances for users with different numbers of articles usually vary substantially, because the latent feature vectors for users are updated along with different directions during the training process. The user topic proportions of tr-MF and tr-CMF serve as an effective constraint to restrict the prediction in certain research areas. This constraint has a strong effect on users with only few ratings. Thus, the performance of tr-MF and tr-CMF on users with only few ratings is expected to be better than that of the competing models. Fig. 7 shows this effect through the performance comparison on users with different numbers of articles. We group the users by the number of the articles in the users' libraries.

As we see from Fig. 7, for users with the number of articles $n \in [20, 40]$ in their libraries, tr-MF outperforms MF with a 3.5% gain on average and CTR with a 2.3% gain on average. While for users with the number of articles less than 20, tr-MF outperforms MF and CTR with a 2% gain. This is due to the fact that the users with the number of articles $n \in [20, 40]$ have more collaborative information than the users with the number of articles less than 20. As is shown, tr-CMF outperforms tr-MF with a 1% gain in almost all cases. It demonstrates that the article relational structure information helps improve the prediction results. In addition, the recalls for users tend to show a decreasing trend with the increase in the number of articles in libraries. The reason for this trend is that more unpopular articles appear in the libraries of those users with a larger number of articles. The predictions for these unpopular articles become difficult.

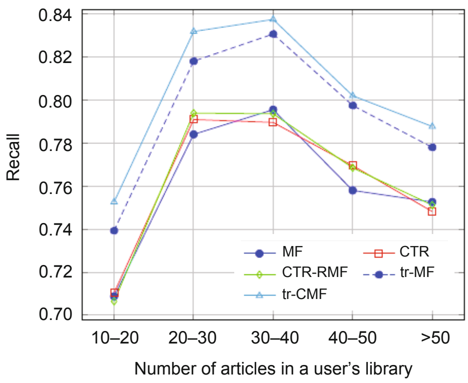

Fig. 7 Comparison of the recalls on different users ($K = 200$, $T = 100$)

6.7 Performance on new users

For new users who have not yet rated any scientific article, the traditional CF algorithms cannot make predictions for these users as these algorithms only use the information about the users' existing ratings. In this section, we demonstrate that the new-user problem can be tackled with the users' activities, such as the browsing histories, the search histories, or the comments.

For this study, since the dataset does not have such new-user samples, we elect to generate such new users synthetically on top of the dataset. 100 users are randomly selected from the dataset as the new users, which are not in the training set. We then randomly select the abstract of a rated article for each new user as his/her searching history. Thus, the corresponding user topic proportions can be generated from the search history for each new user based on the training parameters of tr-MF and tr-CMF. Consequently, we make predictions for these new users with Eq. (14). Since this is the first work addressing the cold-start problem for new users in recommending scientific articles, there is no existing model in the literature that can be used for a comparison directly. We elect to compare tr-MF with a model equivalent to fixing the per-user latent feature vector $\boldsymbol{U}_i = \tilde{\boldsymbol{\theta}}_i$ in the tr-MF model and using per-article latent feature vector $\boldsymbol{V}_j$ to fit to the ratings. This is similar to the method used as a baseline for the cold-start problem in the literature (Wang and Blei, 2011). As this model only uses LDA-like features (topic proportions), we call this comparison method LDA-like.

Fig. 8 illustrates that tr-MF and tr-CMF outperform LDA-like with about 4% and 3% gain on average. The reason for this is that tr-MF and tr-CMF are able to capture the popularities of the articles. Thus, once the user topic proportions are generated from the users' activity records, tr-MF and tr-CMF make accurate predictions for new users. Furthermore, tr-MF behaves better than tr-CMF for the prediction results of new users. The reason for this is that tr-CMF gives a preference to the article relational structures and reduces the impact of topic modeling, which is important for the prediction of new users.

6.8 Recommendation for a specific field

Both tr-MF and tr-CMF obtain two different representations of an article. In this section we take the tr-MF, for example, to illustrate the difference between them; furthermore, we show the comparison study on supporting tasks that are specific to a given field.

6.8.1 Illustration for the two different representations of articles

From the generative process of tr-MF, we see that a scientific article plays two different roles: given a bag-of-word representation, article topic proportions are generated through LDA to regularize the latent feature vectors for users; it serves as the same role of item as other matrix factorization models (Salakhutdinov and Mnih, 2007; Pan *et al.*, 2008). Consequently, we obtain two different representations of article j: $\boldsymbol{\theta}_j$ and $\boldsymbol{V}_j$. First, $\boldsymbol{\theta}_j$ is generated by LDA and represents the article topic proportions. Next, we provide an explanation of $\boldsymbol{V}_j$ in detail. As is seen from the above section, $\boldsymbol{U}_i$ is considered as the topic representation of user i. Specifically, we model the relationship between user i and item j as $\boldsymbol{U}_i^{\mathrm{T}}\boldsymbol{V}_j$. Thus, $\boldsymbol{V}_j$ is represented as the popularity of article j to those topics. Here we denote V_{jk} as the popularity score (p-score) of article j towards the kth topic. Moreover, we find the most applied topics by ranking the p-scores of article j.

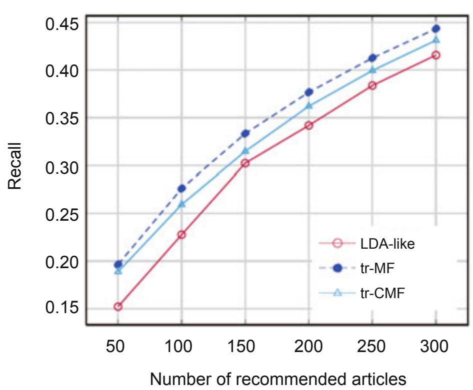

Fig. 8 Comparison of the recalls on the new-user problem ($K = 200$)

We take an example to showcase the difference between $\boldsymbol{\theta}_j$ and $\boldsymbol{V}_j$. Fig. 9 is for the article 'Maximum likelihood from incomplete data via the EM algorithm' (Dempster *et al.*, 1977). The top 5 topics found by $k = \arg\max_k \theta_{jk}$ and $k = \arg\max_k V_{jk}$ are obtained as shown in Figs. 9a and 9b, respectively. Consequently, an interesting fact is discovered, namely that this article is not a statistical biology paper, but researchers engaging in statistical biology typically like this article.

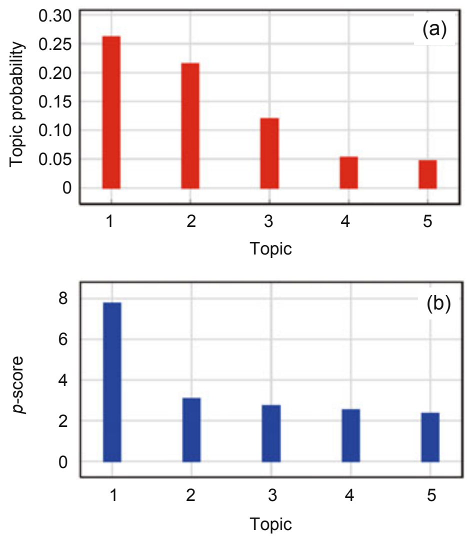

Fig. 9 Maximum likelihood from incomplete data via the EM algorithm: (a) the topic proportions of this article. The top 5 topics are obtained by ranking $\boldsymbol{\theta_{jk}}$ ($\boldsymbol{k = 1, 2, \ldots, K}$). Topic 1: images–analysis–matrix–regression–segmentation; topic 2: hypothesis–testing–maximum–likelihood–bias; topic 3: algorithm–classification–machine–learning–training; topic 4: mathematical–theory–equation–numerical–nonlinear; topic 5: measure–distance–correlation–variance–mean; (b) the popularity score (p-score) distribution on different topics of this article. The top 5 topics are obtained by ranking $\boldsymbol{V_{jk}}$ ($\boldsymbol{k = 1, 2, \ldots, K}$). The item factor entry $\boldsymbol{V_{jk}}$ is denoted as the popularity score of article $\boldsymbol{j}$ towards the $\boldsymbol{k}$th topic. Topic 1: hypothesis–testing–maximum–likelihood–bias; topic 2: networks–models–evolution–sequencing–study; topic 3: biology–statistical–structure–order–research; topic 4: open–partial–frame–closed–generated; topic 5: conservation–missing–frames–predicted–analysis

6.8.2 Recommending papers for a given field using the latent space of articles

Another advantage of tr-MF is to support tasks that are specific to a given field, which is not addressed in the existing literature. Since V_{jk} is considered as the popularity score of article j towards the kth topic, we obtain the most popular articles under this topic found by $j = \arg\max_j V_{jk}, (j = 1, 2, \ldots, M)$. Consequently, tr-MF is expected to more efficiently support tasks that are specific to a given field. Table 2 shows this effect through two example topics with the corresponding top 5 popular articles predicted by tr-MF and CTR, respectively. As is shown, tr-MF is capable of recommending more popular articles for the same topic than CTR demonstrated by both the overall higher Google citations, whereas most existing recommender systems are unable to offer this capability at all.

Since tr-MF is more capable of discovering the popular and fundamental papers for a given topic than CTR or other existing methods, it delivers a better service to researchers to facilitate queries such as 'recommending papers that my paper should cite' and 'recommending papers that may cite my paper'. Specifically, tr-MF is more capable of identifying valuable papers that a researcher wishes to reference but is not familiar with. For example, biological science researchers may wish recommender systems to recommend popular papers on a specific topic in the data mining area that they intend to link their research to, but are not familiar with at all. Here tr-MF offers a better solution.

7 Conclusions

In this paper, we have studied the problem of scientific article recommendation. We note that the interest of users is usually confined to a certain research area. Based on this unique characteristic, we first have developed a new model called tr-MF for recommending scientific articles to users, which is to combine the matrix factorization with a probabilistic topic modeling. Furthermore, to incorporate the relational structure into the framework of tr-MF, we introduce relational matrix factorization. Through combining tr-MF with the relational

Table 2 Comparison of the recommendations for specific fields: the top 5 popular articles under each of two different topics predicted by tr-MF and CTR are listed, respectively

Topic	Model	Paper title	Google citation
W	tr-MF	Optimizing web search using social annotations	324
		Google news personalization scalable online collaborative filtering	264
		The structure of collaborative tagging systems	1687
		Can social bookmarking improve web search?	163
		Usage patterns of collaborative tagging systems	246
			537
	CTR	Tagging and searching: search retrieval effectiveness of folksonomies on the World Wide Web	54
		Personalized interactive faceted search	44
		How are we searching the World Wide Web? A comparison of nine search engine transaction logs	407
		Personalized search based on user search histories	183
		Improving personalized web search using result diversification	78
			153
C	tr-MF	The hallmarks of cancer	13 203
		The cancer genome	360
		The consensus coding sequences of human breast and colorectal cancers	1434
		The epigenomics of cancer	47
		The landscape of somatic copy-number alteration across human cancers	245
			3058
	CTR	The hallmarks of cancer	13 203
		Diverse somatic mutation patterns and pathway alterations in human cancers	96
		Distant metastasis occurs late during the genetic evolution of pancreatic cancer	129
		Genome remodelling in a basallike breast cancer metastasis and xenograft	185
		Integrative molecular concept modeling of prostate cancer progression	287
			2780

Topic W: web-search-user-retrieval-query; topic C: cancer-breast-normal-tumors-growth. The bold numbers are the average values (the data is collected till December, 2011) matrix factorization, we propose the topic regression collective matrix factorization (tr-CMF) model. In addition, we also present the collaborative topic regression model with relational matrix factorization (CTR-RMF) model, which combines the existing CTR model and relational matrix factorization. Compared with the existing literature, the proposed models are particularly effective in recommending articles to users who have only few or even no ratings. Experimental results show that the proposed models perform well against the state-of-the-art methods, such as MF and CTR.

In addition, the proposed models show promising capability for supporting the task of recommending papers in a specific field. Furthermore, it can be applied to many other areas where a bag-of-word representation for items or users is available.

References

Agarwal, D., Chen, B.C., 2009. Regression-based latent factor models. Proc. 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.19-28. [doi:10.1145/1557019.1557029]

Agarwal, D., Chen, B.C., 2010. fLDA: matrix factorization through latent dirichlet allocation. Proc. 3rd ACM Int. Conf. on Web Search and Data Mining, p.91-100. [doi:10.1145/1718487.1718499]

Bell, R.M., Koren, Y., 2007. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. Proc. 7th IEEE Int. Conf. on Data Mining, p.43-52. [doi:10.1109/ICDM.2007.90]

Bell, R.M., Koren, Y., Volinsky, C., 2007. Modeling relationships at multiple scales to improve accuracy of large recommender systems. Proc. 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.95-104. [doi:10.1145/1281192.1281206]

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**:993-1022.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B-Stat. Methdol.*, **39**(1):1-38.

Ge, Y., Liu, Q., Xiong, H., *et al.*, 2011. Cost-aware travel tour recommendation. Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.983-991. [doi:10.1145/2020408.2020568]

Hu, Y., Koren, Y., Volinsky, C., 2008. Collaborative filtering for implicit feedback datasets. Proc. 8th IEEE Int. Conf. on Data Mining, p.263-272. [doi:10.1109/ICDM.2008.22]

Jiang, M., Cui, P., Liu, R., *et al.*, 2012. Social contextual recommendation. Proc. 21st ACM Int. Conf. on Information and Knowledge Management, p.45-54. [doi:10.1145/2396761.2396771]

Jin, X., Zhou, Y., Mobasher, B., 2005. A maximum entropy web recommendation system: combining collaborative and content features. Proc. 11th ACM SIGKDD Int. Conf. on Knowledge Discovery in Data Mining, p.612-617. [doi:10.1145/1081870.1081945]

Koren, Y., 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.426-434. [doi:10.1145/1401890.1401944]

Koren, Y., Bell, R.M., Volinsky, C., 2009. Matrix factorization techniques for recommender systems. *Computer*, **42**(8):30-37. [doi:10.1109/MC.2009.263]

Lafferty, J.D., Blei, D.M., 2005. Correlated topic models. Advances in Neural Information Processing Systems, p.147-154.

Mcauliffe, J.D., Blei, D.M., 2007. Supervised topic models. Advances in Neural Information Processing Systems, p.121-128.

Melville, P., Mooney, R.J., Nagarajan, R., 2002. Content-boosted collaborative filtering for improved recommendations. AAAI, p.187-192.

Pan, R., Zhou, Y., Cao, B., *et al.*, 2008. One-class collaborative filtering. Proc. 8th IEEE Int. Conf. on Data Mining, p.502-511. [doi:10.1109/ICDM.2008.16]

Paterek, A., 2007. Improving regularized singular value decomposition for collaborative filtering. Proc. KDD Cup Workshop, p.5-8.

Purushotham, S., Liu, Y., Kuo, C.C.J., 2012. Collaborative topic regression with social matrix factorization for recommendation systems. Proc. 29th Int. Conf. on Machine Learning, p.759-766.

Salakhutdinov, R., Mnih, A., 2007. Probabilistic matrix factorization. Advances in Neural Information Processing Systems, p.1257-1264.

Salakhutdinov, R., Mnih, A., 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. Proc. 25th Int. Conf. on Machine Learning, p.880-887. [doi:10.1145/1390156.1390267]

Shan, H., Banerjee, A., 2010. Generalized probabilistic matrix factorizations for collaborative filtering. Proc. 10th IEEE Int. Conf. on Data Mining, p.1025-1030. [doi:10.1109/ICDM.2010.116]

Si, L., Jin, R., 2004. Unified filtering by combining collaborative filtering and content-based filtering via mixture model and exponential model. Proc. 13th ACM Int. Conf. on Information and Knowledge Management, p.156-157. [doi:10.1145/1031171.1031201]

Srebro, N., Jaakkola, T., 2003. Weighted low-rank approximations. Proc. 20th Int. Conf. on Machine Learning, p.720-727.

Srebro, N., Rennie, J., Jaakkola, T., 2004. Maximum-margin matrix factorization. Advances in Neural Information Processing Systems, p.1329-1336.

Wang, C., Blei, D.M., 2011. Collaborative topic modeling for recommending scientific articles. Proc. 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.448-456. [doi:10.1145/2020408.2020480]

Weston, J., Wang, C., Weiss, R., *et al.*, 2012. Latent collaborative retrieval. Proc. 29th Int. Conf. on Machine Learning, p.9-16.

Yu, K., Schwaighofer, A., Tresp, V., 2003. Collaborative ensemble learning: combining collaborative and content-based information filtering via hierarchical bayes. Proc. 19th Conf. on Uncertainty in Artificial Intelligence, p.616-623.