Unsupervised Qualitative Scoring for Binary Item Features

Koji Ichikawa¹ · Hiroshi Tamano¹

Received: 14 March 2020 / Revised: 14 May 2020 / Accepted: 25 May 2020 / Published online: 15 June 2020
© The Author(s) 2020

Abstract

Binary features, such as categories, keywords, or tags, are widely used to describe product properties. However, these features are incomplete in that they do not contain several aspects of numerical information. The qualitative score of tags is widely used to describe which product is better in terms of the given property. For example, in a restaurant navigation site, properties such as mood, dishes, and location are given in the form of numerical values, representing the goodness of each aspect. In this paper, we propose a novel approach to estimate the qualitative score from the binary features of products. Based on a natural assumption that an item with a better property is more popular among users who prefer that property, in short, “experts know best,” we introduce both discriminative and generative models with which user preferences and item qualitative scores are inferred from user–item interactions. We constrain the space of the item qualitative score by item binary features so that the score of each item and tag can only have nonzero values when the item has the corresponding tag. This approach contributes to resolving the following difficulties: (1) no supervised data for the score estimation, (2) implicit user purpose, and (3) irrelevant tag contamination. We evaluate our models by using two artificial datasets and two real-world datasets of movie and book ratings. In the experiment, we evaluate the performances of our model under sparse transaction and noisy tag settings by using two artificial datasets. We also evaluate our models’ resolution for irrelevant tags using the real-world dataset of movie ratings and observe that our models outperform a baseline model. Finally, tag rankings obtained from the real-world datasets are compared with a baseline model.

Keywords Label enhancement · Unsupervised learning · Collaborative filtering · Topic model

1 Introduction

Keywords, tags, and categories are widely used to describe product properties. Most e-commerce services, such as Amazon, Alibaba, or eBay, use categories and tags for item filtering. In social bookmarking and recommendation services such as Delicious, Last.fm, and MovieLens, tags are sometimes annotated by users. Even without explicit tags, we unconsciously infer item tags from side information, i.e., product names, explanation texts, reviews, and package designs.

However, such binary expression is incomplete because it does not contain three aspects of numerical information: *quantity*, *relevance*, and *quality*. *Quantity* represents the

strength of a tag with a certain unit. For example, tags, such as low-calorie, light, and hot, lose quantitative information such as 50 kcal, 100 g, and four degrees out of five. With *relevance*, on the other hand, one can see which tag mostly describes the product that does not have a clear unit such as with quantity. For example, if one has a picture of a dog in a park, the tags of that picture may be “dog” and “park.” However, if that picture includes a passerby in the background, it may have the “man” tag. In that case, the relevance of “dog” and “park” should be higher than “man” because the subject of the picture may be “dog in a park.” Finally, *Quality* provides us with information about the goodness of products from the viewpoint of each property. One typical example is the *food*, *location*, and *mood* properties of resort hotels. We often refer to this information in the form of a qualitative score for comparison. We stress the difference between the qualitative score and the other two types of numerical information. Let us consider a protein bar containing high-quality protein. Compared to the quantitative information, the qualitative score is for “which product contains better protein?”

✉ Koji Ichikawa
k_ichikawa0@nec.com

Hiroshi Tamano
h-tamano@nec.com

¹ NEC Corporation, Tokyo, Japan

but not for “how much protein does the product contains?” because the quality of the protein does not directly correspond to the amount of protein. The relevance score is also not directly interpreted as the qualitative score: even for a low-purity protein bar, the tag of “protein” is relevant for that product. This qualitative information is particularly useful for manufacturers and promoters by which they can know how more attractive their product can be to customers compared to their competitors. In this paper, we focus on estimating the qualitative score for each binary item feature.

One difficulty for qualitative score estimation is that we cannot usually obtain supervised data. The qualitative score can be estimated by comparing similar products, i.e., by asking “which hotel is better from the viewpoint of food.” However, a comprehensive survey is unrealistic. More seriously, since scoring is a much more difficult task than tagging, scoring is always affected by psychological biases and human dependence. From this point of view, we focus on a problem setting of unsupervised estimation.

Another difficulty arises when considering an unsupervised setting, because the purpose of the customer is usually implicit in most customer-product interactions such as ratings, purchases, views, clicks, and bookmarks. If user purposes or needs can be determined explicitly, it is possible to obtain a qualitative score by ranking items for each user purpose. One representative example is query and purchase actions on an e-commerce site. A query represents customer demand for products, and a search engine in the site filters products based on the query and lists it. Therefore, by considering the query as an item tag, one can interpret the popularity for each product as the qualitative score for that product in terms of that query. However, in brick-and-mortar shopping, most customers purposes are implicit: we usually do not declare any purpose or query. Even in the e-commerce case, a typical purchase process is that one first filters items by category, query, or stars, reads the descriptions of each item, and finally decides which item to buy. In this process, which point the user prefers is implicit.

We also stress that product features in real-world data often contain tags with low relevance, i.e., tags which do not adequately describe the product. One typical case is tagging for web sites by using keyword extraction. The extraction process always contains a number of irrelevant keywords from links, advertisements, or daily rankings, and one finds only 5–10 keywords that adequately describe the web content. One can also observe many low relevance tagging by user annotation in a social service site. More interestingly, irrelevant tagging occurs when there is a gap between manufacturers and customers, i.e., when tag information provided by manufacturers does not correctly describe item features from the customer’s point of view.

Although label enhancement has recently been addressed, few studies have focused on qualitative score estimation.

In the context of multi-labeled classification, Geng [1] proposed a label enhancement approach for binarized sample labels. Supervised learning for label distribution was developed in the earlier implementation, and a later unsupervised approach was proposed by Shao et al. [2]. They used a similarity assumption between the topology of input features and that of the binary label distributions. However, these approaches are mainly for the item relevance scores discussed above and do not directly compare the label quality between items. On the other hand, item ranking is widely used in the context of a recommendation system that suggests suitable items for specific users [3–14]. However, most studies focused on constructing a personalized item ranking for each user and did not provide information about the qualitative aspect of the apparent item binary features.

In this paper, we propose a novel approach for qualitative score estimation from binary and noisy item features, which contributes to resolving the difficulties mentioned above, i.e., (1) no supervised data, (2) implicit customer purpose, and (3) noisy tags. With this approach, we infer the qualitative score based on the assumption that an item with a better property is more popular among users who prefer that property, in short, “experts know best.” One discriminative and two generative models are introduced with this approach with which user preferences and item qualitative scores are inferred from user-item interactions. One novelty is that we constrain the space of the item qualitative score by the item binary features so that it can only have a numerical value according to their tags. We evaluate our models by using two artificial and two real-world datasets. In the experiment of the two artificial datasets, we evaluate the performances of our model under the sparse transaction and noisy tag settings. We also evaluated our models’ resolution for irrelevant tags using the real-world dataset of movie ratings and observed that our models outperformed a baseline model. Tag rankings obtained with our models from the two real-world datasets were compared with a baseline ranking model.

The organization of this paper is as follows: In Sect. 2, we review recent related work. In Sect. 3, we discuss the proposed approach and models. In Sect. 4, we present the minimization and inferences for our models. We present the experimental results in Sect. 5 and finally give a discussion and conclusion in Sect. 6.

2 Related Work

We infer both qualitative tag score and user preferences by using user-item interactions such as ratings, purchases, views, clicks, and bookmarks. Collaborative filtering is a conventional approach for latent feature extraction from user-item interactions. With this approach, the user and

item latent features are extracted through low-dimensional matrix factorization for user--item interactions. Earlier studies focused on data with user ratings [3–12], and later studies developed approaches for implicit data such as clicks, bookmarks, and purchases [13, 14]. Probabilistic approaches for feature extraction were also proposed by using the framework of the topic model [15–17]. The topic model was first introduced as latent semantic analysis (LSA) [15] for document clustering and probabilistic extension was later proposed [16]. Latent Dirichlet allocation (LDA) [17] is a Bayesian extension of the probabilistic LSA in which a document has topic distribution, and terms of the document are generated by the term distributions of the topics. They give interpretable latent topics, giving the word distribution of each topic. Many extensions have been proposed in terms of document relation [18], topic correlation [19], hierarchical structure of topics [20, 21], and combining meaning of words [22–27]. In collaborative LDA, Wang and Blei [28] proposed a combined model of the topic model for item documents and matrix factorization for user--item interactions. One significant difference from our models is that their model uses the rich text information of items, such as articles or reviews, and constructs the topic distribution for item features. Our approach, however, uses sparse, noisy binary tags for item features.

As we mentioned in introduction, the qualitative score relates to tag relevance. In the context of multi-label classification, label importance was introduced by Geng [1], with which label distribution is inferred with supervised learning. Later, Shao et al. [2] proposed an unsupervised approach, with which numerical label distribution is inferred under a constraint of given binary labels, assuming the similarity between the topology of input features and that of the label distributions. Tag feature extraction was also investigated in the context of tag recommendation, with which proper tags are recommended to annotators. The latent feature of the tag was constructed from a tensor decomposition of user--item--tag relations, enabling personalized tag recommendation [29, 30].

3 Model Descriptions

In this section, we introduce our tag scoring models, where the qualitative tag score is inferred from user--item interactions and binary item tags. We first give notations and the problem, define the qualitative tag score, and finally describe our models. We propose three models. One is based on matrix factorization, and the other two are probabilistic realizations based on the topic model.

Table 1 Example of binary features

	Hot	Curry	Chicken
Chili Chicken	1	0	1
Chicken Curry	1	1	1
Beef Curry	1	1	0

Table 2 Example of probabilistic qualitative scores

	Hot	Curry	Chicken
Chili Chicken	0.3	0	0.7
Chicken Curry	0.3	0.4	0.3
Beef Curry	0.4	0.6	0

3.1 Problem Setting and Notations

We consider user--item interactions composed of a set of users $\mathcal{U} = \{u_1, u_2, \dots, u_U\}$ and items $\mathcal{I} = \{i_1, i_2, \dots, i_I\}$. Here, U and I represent the number of users and items, respectively. We represent user--item interactions by a form of an adjacency matrix. For countable interactions, such as purchases and clicks, we consider $Y_{u,i} \in \mathbb{N}^{U \times I}$ where $Y_{u,i} = n$ if user $u \in \mathcal{U}$ interacts with item $i \in \mathcal{I}$ for n times. For interactions with numerical values such as ratings, the adjacency matrix can be represented as $Y_{u,i} \in \mathbb{R}^{U \times I}$ where $Y_{u,i} = r$ if user $u \in \mathcal{U}$ gives rating values of r to item $i \in \mathcal{I}$. In both cases, $Y_{u,i} = 0$ implies a missing sample. For simplicity, we consider user purchases of items. We also represent the set of items purchased by user u by $\mathcal{Y}_u = \{i_{u1}, i_{u2}, \dots\}$ and the set of all user--item interactions by $\mathcal{Y} = \{(u, i) \mid i \in \mathcal{Y}_u\}$. We define a set of tags $\mathcal{K} = \{k_1, k_2, \dots, k_K\}$ and express item features by a binary matrix of tags $T \in \{0, 1\}^{K \times I}$. In our setting, we assume an item has several tags and $T_{k,i} = 1$ if item $i \in \mathcal{I}$ has tag $k \in \mathcal{K}$.

The problem is to estimate keyword score $\Phi \in \mathbb{R}^{K \times I}$ under observed Y and T , where $\Phi_{k,i}$ represents the qualitative score of an item i in terms of a tag k . Simple examples are given in Tables 1 and 2, where item binary feature T and qualitative score Φ correspond to Tables 1 and 2, respectively.

3.2 Definition of Qualitative Score and Model Concept

To clarify the model concept, we first define qualitative score for our approach. Our definition of the qualitative score is that, in a nutshell, *if an item has a tag with higher qualitative score, then that item is more popular or attractive for users who prefer that property*. In the following, we will explain this concept in more detail and introduce two models derived from the definition.

As we mentioned in introduction, the simplest definition can be obtained from manual scoring. However, since manual scoring is difficult and affected by psychological bias, scoring based on user-item interactions is preferable. One simple qualitative score is obtained using the popularity of items: by comparing item popularity tag by tag, we can construct tag rankings. However, this scoring gives the same tag score for one product. For example, movies with fantasy and sci-fi tags have the same score for these two tags. We propose a user preference based score where an item with a higher tag score implies it is more popular or attractive for users who prefer that property. From this definition, the behavior of experts of each tag determines each tag score; therefore, we describe this concept as the “experts know best” assumption. In the previous example, if a movie is more popular among those who like fantasy, a higher score is assigned to the fantasy tag than the sci-fi tag. In other words, we assume that item purchase can be approximated by a combination of user preference and tag scoring. This definition naturally enables the derivation of an extension of matrix factorization. In this paper, we call this model factorized feature scoring model (FFSM).

We can also define the score by a probability. In this definition, $\Phi_{k,i}$ represents a probability that item i is chosen in terms of tag k . Table 2 illustrates a simple example. We have three items that have tags of *hot*, *curry*, and *healthy*, and each column is normalized to 1 for the probabilistic definition. The score shows that when one wants something *hot*, Chili Chicken, Chicken Curry, and Beef Curry are chosen with probabilities of 0.3, 0.3, and 0.4. Similarly, when one wants *curry*, Chicken Curry and Beef Curry are chosen with a probability of 0.4 and 0.6. Since Chili Chicken is not a curry, i.e., does not have the *curry* tag, the probability is zero. By using this definition, we model the purchase by extending the topic model, which we call the probabilistic feature scoring model (PFSM). Note that in this definition, the item selection probability of a given tag is inferred from the selection behavior of users who prefer that tag; therefore, it is also based on the “experts know best” assumption.

3.3 Factorized Feature Scoring Model

Based on the first definition of the qualitative score, we develop a matrix factorization model. The basic idea of this model is that matrix Y can be approximated by decomposing low-rank matrices P and Q : $Y_{ui} \sim \sum_{d=1}^D P_{ud}(Q^T)_{di}$, where we define $P \in \mathbb{R}^{M \times D}$, $Q \in \mathbb{R}^{N \times D}$, and $D \ll U, I$. The rows of P represent the user’s latent preferences, and the rows of Q represent the item’s latent features. The notation D is the dimension of the latent space. The root mean squared error (RMSE) between Y and PQ^T is minimized under the regularization of P and Q , which derives the object function L as

$$L = \sum_{u,i} w_{ui} \left(Y_{ui} - \sum_d P_{ud} Q_{id} \right)^2 + \lambda \left(\sum_{u,d} P_{ud}^2 + \sum_{i,d} Q_{id}^2 \right),$$

where $p_{ud} (q_{id}) \in \mathbb{R}$ denotes the entry of $U (V)$ in the $u(i)$ th row and d th column, and the hyper-parameter λ controls the strength of the regularization. Weight $w_{ui} \in [0, 1]$ is introduced for missing samples, which takes $w_{ui} = 1$ if user u purchases item i (i.e., $Y_{ui} > 0$), otherwise a small value. We simply adopt the uniform weight [13].

With FFSM, we define the tag-constrained objective function by assuming a product of user preference Θ and the qualitative tag score Φ explains the purchase matrix Y :

$$L = \sum_{u,i} w_{ui} \left(Y_{ui} - \sum_k \Theta_{u,k} \Phi_{k,i} \right)^2 + \lambda \sum_{u,k} \Theta_{u,k}^2 + \lambda \sum_{k,i} \text{Reg}(\Phi_{k,i}, T). \quad (1)$$

In our approach, we minimize L under the nonnegative constraint: $\Theta_{u,k}, \Phi_{k,i} > 0$. In contrast to the original matrix factorization, we let the dimension of the latent space D be equal to the dimension of the tag space, i.e., $D = K$ and introduce the tag constraint Reg :

$$\text{Reg}(\Phi_{k,i}, T) = \begin{cases} \Phi_{k,i}^2 & (T_{k,i} = 1) \\ c_\infty \Phi_{k,i}^2 & (T_{k,i} = 0) \end{cases}. \quad (2)$$

Here c_∞ is a large constant that suppresses Φ at $T_{ik} = 0$. In this paper, we adopt the hard constraint, where $c_\infty = \infty$, i.e., $\Phi_{k,i} = 0$ if item i does not have tag k .

3.4 Probabilistic Feature Scoring Model

We develop this model based on the topic model. In the topic model, the purchase history is modeled by a product of the probability of which user preference appears at this purchase and the item distribution with respect to the chosen preference:

$$P(Y_{u,i} = 1 | \Theta, \Phi) = \sum_z P(Y_{u,i} = 1 | z, \Phi) P(z | \Theta). \quad (3)$$

here $\Theta \in \mathbb{R}^{U \times D}$ and $\Phi \in \mathbb{R}^{D \times I}$ and D is the dimension of the topics. The Θ and Φ rows represent distributions of a user’s latent preferences and item’s latent topics, respectively. Topic index $z \in \{1, 2, 3, \dots, D\}$ can be interpreted as the purpose of user u for purchase i and is sampled from the multinomial distribution of the user preference Θ : $P(z | \Theta) = \Theta_{u,z}$, and item i is chosen according to a multinomial distribution of $\Phi_{k,i}$: $P(Y_{u,i} = 1 | z, \Phi) = \Phi_{z,i}$.

In this model, we define the matrices of Θ and Φ based on the tag space T . First, we let D be equal to the dimension of

Fig. 1 Graphical model of PFSM

the tag space, i.e., $D = K$. Moreover, we impose a constraint where Φ is nonzero if and only if $T \neq 0$, i.e.,

$$\Phi_{k,i} \begin{cases} > 0 & (T_{k,i} = 1) \\ = 0 & (T_{k,i} = 0) \end{cases}, \quad (4)$$

satisfying $\sum_i \Phi_{k,i} = 1$. By introducing the constraint from the keyword space, we can interpret that each component of $\Phi_{k,i}$ is the qualitative score defined in the previous subsection.

The posterior probability $\prod_{u,i \in \mathcal{Y}} P(\Theta, \Phi | Y_{u,i})$ is inferred by introducing priors of Θ and Φ . For Θ , we adopt the conjugate prior for the multinomial distribution:

$$P(\Theta) = \text{Dir}(\Theta | \alpha), \quad (5)$$

where Dir is the Dirichlet distribution and $\alpha \in \mathbb{R}^{U \times K}$ is a hyperparameter of Dir . Similarly, we introduce the Dirichlet prior for Φ . In this case, since Φ is constrained by T , the sample space of the prior is also constrained. We denote

$$P(\Phi_k) = \text{Dir}(\Phi_k | \beta, T), \quad (6)$$

and a hyperparameter $\beta \in \mathbb{R}^{K \times I}$ takes $\beta_{k,i} = 0$ when $T_{k,i} = 0$. Figure 1 illustrates the graphical model of the generative process.

3.5 Correlated Probabilistic Feature Scoring Model

We extend PFSM by incorporating the tag--tag correlation. For user preference, we can naturally consider that preferences with a similar meaning are correlated. For example, one with a higher preference for “humor” tends to have a higher value for “comedy.” Since each dimension k of the topic matrix Θ can connect to a tag name, we use the external knowledge for topic correlation. On the other hand, we conjecture that the correlation from meaning of words is not always valid for user preference. Therefore, we use the word correlation matrix as a

prior of this extended model, which we call correlated PFSM (CPFSM) and infer topic correlation with it.

Based on the correlated topic model [19], we introduce a multivariate normal distribution for correlated topics. The generative process of CPFSM can be described as follows:

- Draw the means and covariances of user topics: $\mu, \Sigma \sim \text{NIW}(\mu_0, \Sigma_0, \lambda, \nu)$. Here NIW is the normal-inverse-Wishart distribution, which is a conjugate prior for the multivariate normal distribution. Prior knowledge μ_0, Σ_0 control mean and covariance matrix for a topic parameter, respectively, and λ and ν are hyperparameters of the distribution.
- Draw a parameter for user topic distribution η by a multivariate normal distribution: $\eta \sim \text{N}(\mu, \Sigma)$. Here $\text{N}(\mu, \Sigma)$ represents a normal distribution with a mean of μ and covariance matrix of Σ .
- Convert η to Θ by the softmax function: $\Theta_{uk} = e^{\eta_{uk}} / \sum_k e^{\eta_{uk}}$.
- Draw item distribution Φ by a constrained multinomial distribution in Eq. (4).
- For each purchase with index c of user u , draw user attention $z_{u,c}$ from the multinomial distribution of Θ : $z_{u,c} \sim \text{Mult}(\Theta)$.
- Draw user select item $i_{u,c}$ from the multinomial distribution of $\Phi_{z_{u,c}}$: $i_{u,c} \sim \text{Mult}(\Phi_{z_{u,c}})$.

Similar to PFSM, we impose a constraint where Φ is nonzero if and only if $T \neq 0$. Figure 2 illustrates the graphical model of the generative process.

4 Minimization and Inferences

4.1 Minimization for FFSM

For the minimization of Eq. (1), we develop an efficient algorithm for FFSM based on the alternative least square approach proposed by [31]. In the algorithm, we carry out element-wise minimizations of $\Theta_{u,k}$ and $\Phi_{k,i}$. The update equation for $\Theta_{u,k}$ is given by

$$\Theta_{u,k} = \frac{\sum_i w_{u,i} \Phi_{k,i} Y_{u,i} - \sum_i \sum_{k' \neq k} w_{u,i} \Phi_{k,i} \Theta_{u,k'} \Phi_{k',i}}{\lambda + \sum_i w_{u,i} \Phi_{k,i}^2}, \quad (7)$$

for any $1 \leq u \leq U$ and $1 \leq k \leq K$. Similarly, the update equation for $\Phi_{k,i}$ is given by

$$\Phi_{k,i} = \frac{\sum_u w_{u,i} \Theta_{u,k} Y_{u,i} - \sum_u \sum_{k' \neq k} w_{u,i} \Theta_{u,k} \Phi_{k',i} \Theta_{u,k'}}{\lambda + \sum_u w_{u,i} \Theta_{u,k}^2}, \quad (8)$$

Fig. 2 Graphical model of CPFSM

if $T_{k,i} = 1$ and otherwise $\Phi_{k,i} = 0$ for any $1 \leq i \leq I$ and $1 \leq k \leq K$. To satisfy the nonnegative constraint for Θ and Φ , we set $\Theta = 0$ and $\Phi = 0$ for negative values for each iteration.

Algorithm 1 illustrates the pseudo-code for this algorithm.

4.2 Inference for PFSM

In PFSM, most of the inference techniques developed for vanilla LDA can be applied with a small modification, i.e., by constraining the word distribution of Φ according to T . We adopt the collapsed Gibbs sampling, where $z_{u,c}$ is sampled after marginalizing Φ and Θ :

$$P(z_{u,c} = k | z^{\setminus u,c}, Y, \alpha, \beta) \propto \left(\frac{\beta_{k,i_{u,c}} + n_{k,i_{u,c} \setminus u,c}}{\sum_i \beta_{k,i} + n_{k,i \setminus u,c}} \right) \left(\frac{\alpha_{u,k} + n_{u,k \setminus u,c}}{\sum_{k'} \alpha_{u,k'} + n_{u,k' \setminus u,c}} \right). \quad (9)$$

for $T_{ik} = 1$ and 0 otherwise. We define that $n_{k,i}$ is the total number of topics k assigned to an item i and $n_{u,k}$ is the total number of topics k appearing in purchases of a user u :

$$\begin{aligned} n_{k,i} &= \sum_{u,c} I(z_{u,c} = k) I(i_{u,c} = i), \\ n_{u,k} &= \sum_c I(z_{u,c} = k), \end{aligned} \quad (10)$$

and $\cdot \setminus u, c$ represents the calculation excluding the c th purchased item by user u .

The elements of Φ and Θ can be written as:

$$\Phi_{ki} \propto \begin{cases} \beta_{k,i} + n_{k,i} & (T_{k,i} = 1) \\ 0 & (T_{k,i} = 0) \end{cases}, \quad (11)$$

and

$$\Theta_{uk} \propto \alpha_{u,k} + n_{u,k}.$$

Algorithm 1 FFSM

```

inputs  $w, \lambda, Y, T$ .
initialize  $\Theta, \Phi$ .
repeat
  for  $u = 1, \dots, U$  do
    for  $k = 1, \dots, K$  do
      Update  $\Theta_{u,k}$  by Eq. (7).
       $\Theta_{u,k} \leftarrow 0$  if  $\Theta_{u,k} < 0$ .
    end for
  end for
  for  $i = 1, \dots, I$  do
    for  $k = 1, \dots, K$  do
      Update  $\Phi_{i,k}$  by Eq. (8).
       $\Phi_{i,k} \leftarrow 0$  if  $\Phi_{i,k} < 0$ .
    end for
  end for
until convergence

```

Algorithm 2 PFSM

```

inputs  $\alpha, \beta, Y, T$ .
initialize  $\Theta, \Phi$ , and  $z$ , and compute  $n_{u,k}$  and  $n_{k,i}$  by Eq. (10).
repeat
  for  $u = 1, \dots, U$  do
    for  $c \in \mathcal{Y}_u$  do
       $z \leftarrow z_{u,c}$ 
       $n_{u,z} \leftarrow n_{u,z} - 1$ 
       $n_{z,i} \leftarrow n_{z,i} - 1$ 
      Sample  $z$  by Eq. (9).
       $n_{u,z} \leftarrow n_{u,z} + 1$ 
       $n_{z,i} \leftarrow n_{z,i} + 1$ 
       $z_{u,c} \leftarrow z$ 
    end for
  end for
until convergence

```

Algorithm 3 CPFSM

inputs $\Sigma_0, \lambda, \nu, \mu_0, \beta, Y, T$.
initialize Θ, Φ , and z , and compute $n_{u,k}$ and $n_{k,i}$ by Eq. (10).
repeat
 for $u = 1, \dots, U$ **do**
 for $c \in \mathcal{Y}_u$ **do**
 $z \leftarrow z_{u,c}$
 $n_{u,z} \leftarrow n_{u,z} - 1$
 $n_{z,i} \leftarrow n_{z,i} - 1$
 Sample z by Eq. (12).
 $n_{u,z} \leftarrow n_{u,z} + 1$
 $n_{z,i} \leftarrow n_{z,i} + 1$
 $z_{u,c} \leftarrow z$
 end for
 for $k = 1, \dots, K$ **do**
 Sample η by Eq. (13).
 Sample λ by a normal distribution function with the mean and standard deviation of $PG(n_u, \eta_{u,k} - \zeta_{u,k})$.
 end for
 Update Σ and μ by Eq. (14).
 end for
until convergence

Algorithm 2 illustrates the pseudo-code for this algorithm.

4.3 Inference for CPFSM

For CPFSM, we also infer the parameter of the model by using the Gibbs sampling algorithm. Although $P(\eta)$ is no longer a conjugate distribution to $P(z|\eta)$, an efficient sampling algorithm for z and η has been proposed by [32]. The sampling probability of $z_{u,c} = k$ is given by the Φ -collapsed posterior function:

$$P(z_{u,c} = k | z^{\setminus u,c}, Y, \eta, \beta, T) \propto \left(\frac{\beta_{k,i_{u,c}} + n_{k,i_{u,c} \setminus u,c}}{\sum_i \beta_{k,i} + n_{k,i \setminus u,c}} \right) \left(\frac{e^{\eta_{u,k}}}{\sum_{k'} e^{\eta_{u,k'}}} \right). \quad (12)$$

and we obtain the tag score by Eq. (11). On the other hand, the scale mixture representation of $p(\eta_{u,k} | \eta^{\setminus k})$ enables integration of auxiliary variable $\lambda_{u,k}$ (see the study by Chen et al. [32] for more detail). By utilizing this technique, we finally obtain the sampling distribution for η and λ :

$$P(\eta_{u,k} | \eta^{\setminus k}, z, \Sigma) \propto N(\gamma_{u,k}, \tau_{u,k}), \quad (13)$$

where we define

$$\gamma_{u,k} = \tau_{u,k}^2 (\mu_{k,u} / \Sigma_{k,k} + n_{u,k} - n_u / 2 + \lambda_{u,k} \zeta_{u,k}),$$

$$\tau_{u,k} = (1 / \Sigma_{k,k} + \lambda_{u,k})^{-1/2}.$$

here $n_u = \sum_k n_{u,k}$, $\mu_{k,u} = \mu_k - \Sigma_{k,k} \Lambda_{k,\setminus k} (\eta_{u,\setminus k} - \mu_{\setminus k})$, and $\zeta_{u,k} = \ln(\sum_{l \neq k} e^{\eta_{u,l}})$. The sampling distribution for $\lambda_{u,k}$ is given by $PG(n_u, \eta_{u,k} - \zeta_{u,k})$ where PG is the Polya--Gamma distribution.

Since n_u is generally large in our case, we can approximate PG to a normal distribution function with the mean and standard deviation of $PG(n_u, \eta_{u,k} - \zeta_{u,k})$.

For μ and Σ , due to the conjugacy between NIW and multivariate normal distributions, posterior distribution can be parameterized by NIW: $P(\mu, \Sigma | \eta) \propto P(\eta | \mu, \Sigma) P(\mu, \Sigma) = NIW(\hat{\mu}, \hat{\Sigma}, \hat{\lambda}, \hat{\nu})$, where we define $\hat{\mu} = (\lambda \mu_0 + U \bar{\eta}) / (\lambda + U)$, $\hat{\Sigma} = \Sigma_0 + \sum_u (\eta_u - \bar{\eta})(\eta_u - \bar{\eta})^T + \frac{\lambda U}{\lambda + U} (\bar{\eta} - \mu_0)(\bar{\eta} - \mu_0)^T$, $\hat{\lambda} = \lambda + U$, and $\hat{\nu} = \nu + U$. Here, $\bar{\eta}$ is the mean of η : $\bar{\eta} \equiv 1/U \sum_u \eta_u$. The MAP inference gives

$$\begin{aligned} \Sigma_{MAP} &= \hat{\Sigma} / (\hat{\nu} + K + 1), \\ \mu_{MAP} &= \hat{\mu}. \end{aligned} \quad (14)$$

Algorithm 3 illustrates the pseudo-code for this algorithm.

5 Consistency of Tag Score

We now discuss the consistency of the rankings. In our models, we approximate Y by non-negative Φ and Θ , which can be interpreted as one realization of the non-negative matrix factorization (NMF). Generally, NMF does not always provide the same solution for Φ and Θ . As a result, the obtained tag score may differ trial by trial. The reason for this inconsistency stems from the following two reasons: non-uniqueness and local minimum.

Non-uniqueness of non-negative matrix factorization is a well-known problem. Even after removing trivial degrees of freedom, such as permutation of the components or scaling, the solution of NMF is not generally unique. The uniqueness relates to the form of the decomposition, and the sufficient and necessary condition for the uniqueness has been investigated [33–36]. In our model, thanks to the binary tag constraints, the decomposition tends to offer the unique solution: The scaling arbitrariness is removed by conditioning $\sum_k \Phi_{k,i} = 1$, and one can easily see that the permutation is not allowed unless there exist a pair k and $k' \neq k$ that satisfies $T_{k,i} = T_{k',i}$ for all $i \in \mathcal{I}$. Moreover, a recent study has observed that if T and transaction Y are sparse, the decomposition tends to offer a unique solution [36].

Another possibility of this inconsistency comes from the fact that global minimum search is NP hard [37] and there exist local minimums where another list of tag scoring would be suggested. One remedy for this local minimum is to run an algorithm several times by changing the random seeds and observe the behavior of the list. In this paper, we

simply discard tags with unstable rankings and leave further development for future work.

6 Experiments

6.1 Model Performance Evaluation

We first evaluated our models' performance by using two artificial datasets.

6.1.1 Dataset

We generated two artificial datasets. In the first dataset, we assumed that user-item interactions were generated in a discriminant model fashion, i.e., via the product of user preferences and item tag scores. In this dataset, we randomly generated Φ and Θ in a range of 0 to 1 and normalized them along the tag direction and the user direction, respectively. The user-item interaction Y was constructed from the product of Φ and Θ . By imposing thresholds on Φ and Y , we binarized them. The threshold values were determined by the given number of tags and user purchases, respectively. Below, we call this dataset **D1**. In the second dataset, we assumed that the user-item interactions were generated via the generative process described in Sect. 3. We first generated an item binary tag matrix T under a fixed number of tags for each item (we adopted 5 and 15) then generated Φ and Θ by using Eqs. (5) and (6), respectively, under the constraint of T , setting $\alpha, \beta = 0.5$, and finally generated Y by using Eq. (3). Below, we call this dataset **D2**. We fixed the user, item, and keyword spaces to $U = 500$, $I = 500$ and $K = 50$ for both models. We generated these two datasets ten times and computed the average scores and standard deviations.

6.1.2 Evaluation Metrics

We evaluated the models' performance by introducing two standard ranking metrics:

- *HitRate (HR)* When evaluating the qualitative score, it is often the case that the ranking of items is more important than the numerical score. To evaluate qualitative ranking performance, we introduce HR. The HR for top N items

(HR@N) is an intuitive ranking metric with which sets of the top N items of the prediction and the ground truth are compared. The HR provides the ratio of the intersection of the two item sets:

$$\text{HR@N} = \frac{1}{K} \sum_k \frac{1}{N} |\{\text{top } N \text{ of } \Phi_k\} \cap \{\text{top } N \text{ of } \hat{\Phi}_k\}|. \quad (15)$$

- *Normalized discounted cumulated gain (nDCG)* The second metric is for considering both ranking and score accuracy. Intuitively, score accuracy for higher ranking items is more important than that for lower ranking items. DCG is a popular weighted rank metric with which the larger weight is imposed on higher-ranked items. DCG@N for tag k is defined by

$$\text{DCG}_k@N = \sum_{n=1}^N \frac{2^{\text{rel}(k, i_n^k)} - 1}{\log_2(n+1)}. \quad (16)$$

We assume that a model predicts the top N items of tag k : $\{i_1^k, i_2^k, \dots, i_N^k\}$ and $\text{rel}(k, i)$ represents the relevance of item i for tag k . In our problem setting, the relevance is $\text{rel}(k, i) = \Phi_{k,i}$. The nDCG is defined by $\text{nDCG}_k@N = \text{DCG}_k@N / \text{IDCG}_k@N$, where IDCG represents the ideal DCG, in which the top k item index series $i_1'', i_2'', \dots, i_k''$ are computed from the ground truth $\Phi_{k,i}$. For the metric of our experiments, we used the averaged nDCG: $\text{nDCG@N} = \frac{1}{K} \sum_k \text{nDCG}_k@N$.

6.1.3 Experimental Setting

The first experiment was for observing the relationship between the accuracy and sparsity of a dataset. The transaction data were generated by changing the number of interactions in a range from 5 to 100 for each user. Performance was measured using the metrics introduced above. For the second experiment, we evaluated the models performance under contaminated tag features. We generated Φ by the original T and added noise tags. The noise tags were generated by a probability of tag frequency. By using the contaminated tag feature, we fitted the transaction. We changed the number of contaminations from 1 to 30 for each item and obtained the metric scores. In both experiments, we generated a validation dataset for each number of interactions/contaminations and tuned the model hyperparameters by Bayesian search [38, 39]. The hyperparameters were determined from the results of the 50 trials. We set the search range $w \in [0.001, 1]$, and $\lambda \in [0.001, 100]$ for FFSM and $\alpha, \beta \in [0.0001, 10]$ for PFSM.

Fig. 3 Performance regarding sparsity

Fig. 4 Model performance regarding tag contamination

6.1.4 Baseline Model

We adopted a random prediction- and a popularity-based scoring model as the baseline.¹ For each tag k , we define item set $\mathcal{I}_k$ in which items have tag k . We computed the qualitative score for item i in $\mathcal{I}_k$ by accumulating user interactions to that item. The formula for the popularity score can be written as

$$S_{ik}^{\text{pop}} = \frac{\sum_u Y_{u,i} T_{ik}}{\sum_{u,i} Y_{u,i} T_{ik}}. \quad (17)$$

¹ Since unsupervised qualitative score estimation is rather new concept, there are few existing methods. We investigated the performance of the top-n ranking method, such as the weighted regularized matrix factorization [13, 14] and Bayesian personalized ranking [40], which utilize the structure of the tag matrix to predict the scores. However, we confirmed that the prediction accuracy was not significantly better than the random prediction because the tag matrix was generated randomly.

6.1.5 Results

Figures 3 and 4 show the results. In both experiments, we observed that our models significantly outperformed the baseline models over a wide range of the number of interactions/contaminations. The performance decreased as the number of tags increased, which can be seen by comparing the scores of #Tags = 5 and #Tags = 15 as well as from the results of the second experiments. This implies that the tag score is more difficult to resolve in a dense tag situation. Comparing the performances of FFMS and PFSM, we found that the FFMS gave competitive performances even in the generative dataset (D2).

6.2 Performance Evaluating Application to Tag Relevance

In this experiment, we evaluated our models performance using a real-world dataset. Since our feature scoring is a new problem setting, few open datasets are available regarding user behavior and item qualitative score. Therefore, we

Fig. 5 Model performances regarding relevance score

instead evaluated our models performance regarding relevance scoring, in which we assumed tags with low qualitative scores are likely to be irrelevant.

6.2.1 Dataset

We adopted the MovieLens dataset.² The dataset has transaction data of user ratings to movies, which consists of 610 users, 9724 items, and 100,836 ratings. We extracted samples rated above 3.5 as positive samples. For the relevance score of movies, we used the tag genomes [41], for which the relevance scores of 1128 tags were accumulated from questionnaires for a small sample of movies and a prediction model was constructed to complete the relevance scores for other movies. We aimed to demonstrate how our models precisely reconstruct the relevance score without supervised data using only user--item interactions and contaminated binary tag information.

6.2.2 Experimental Setting

We evaluated the performance of FFSM, PFSM, and predictions combined with the tag frequency. The last predictions are obtained by the product of the tag frequency and the tag scores of FFSM and PFSM. We constructed a binary tag matrix by imposing thresholds so that each movie had the top ten relevant tags and binarized these top ten relevant tags. We considered these tags as relevant tags. We selected items that appeared at least ten times in the transaction. To evaluate our models resolution, we added artificial irrelevant tags to the original binary feature. The irrelevant tags were generated by the probability of tag frequency. We changed the number of contaminated tags and evaluated model

resolution by using HR and nDCG. The metric scores were computed for each item not for each tag.

6.2.3 Baseline Model

We adopted the tag frequency based the relevance score for the baseline model. Based on an intuitive concept in which a less frequent tag is likely less relevant, we define the score of tag k for item i as $S_{ik}^{freq} = \sum_i T_{i,k} \cdot s$

6.2.4 Results

Figure 5 shows the results. Both FFSM and PFSM resulted in higher scores than the baseline in a region of low contamination, while the tag frequency weighted prediction of FFSM significantly outperformed the other models above the tag contamination of 50 percent. Since our model was not optimized for the irrelevant tag removal task, other methods that focus on that task may produce better results.³ We note, however, that combining our method with other approaches may yield better accuracy.⁴

² <https://grouplens.org/datasets/movielens/>.

³ We compared our models to the weighted regularized matrix factorization (WRMF) [13, 14] that decomposes the tag matrix into the product of two low-rank matrices. Even though WRMF is a strong baseline algorithm in the top-n ranking task, we found that the performances of our model were competitive.

⁴ For example, we attempted to perform WRMF after setting the tags in the lower 10 percent of the FFSM score to 0. As a result, we found that this preprocessing improved the accuracy by about 3–5 percent. More rigorous experiments and evaluation against a broader range of models are left to future research.

6.3 Exploration for the Qualitative Scores

Finally, we show the item ranking for each tag obtained with our models using two real-world datasets.

6.3.1 Dataset

We adopted Moivelens dataset and Goodbooks-10k dataset [42], where 10,000 user's ratings to 53,424 books were provided. In the Goodbooks dataset, the number of user-annotated tags is also provided, and we used them for the item features.

6.3.2 Experimental Setting

We implemented CPFSM and compared the results with those from the baseline ranking models. For the baseline ranking, we adopted the popularity based ranking introduced in Eq. (17) for MovieLens and used the tag count for Goodbooks. For the binary item features, we used the genre tags for MovieLens and the tag count for Goodbooks. In the Goodbooks dataset, we made a set of tags shown in Fig. 6 by merging tags with orthographic variants or similar meaning and filtered the books having at least one tag among them. We considered a series of books as the same title. We filtered the transaction so that it contains books with at least > 10 times ratings of > 3.5 and users with > 100 interactions. We constructed the prior correlation matrix Σ_0 by the cosine similarity of the word embeddings of FastText [43]. To obtain Σ similar to the word relation, we scaled the word similarity matrix S to $\Sigma_0 = \lambda(\nu - K - 1)S$. We fixed $\mu_0 = 0$, $\nu = U + K$ and $\lambda = U$.

6.3.3 Results

We show some of the tag rankings in 3 and 4. With the baseline, the titles with high popularity were highly ranked for different tags. For example, the Star Wars series was ranked at the top for both sci-fi and adventure tags in MovieLens. Due to the decomposition of the user preferences, CPFSM provided other rankings for both tags: Interstellar, Inception, and Arrival for sci-fi and the Lord of the Ring series for adventure. Similarly, while Harry Potter was the highest ranked for adventure, fantasy, and children tags by tag count ranking in the Goodbooks dataset, CPFSM provided more divergent rankings, in which experts of each tag may prefer these books.

We also observed the keyword correlation obtained from correlated LDA. Figure 6 shows a word relation graph of the Goodbooks dataset by imposing a threshold to the correlation matrix. The difference between the original prior and posterior correlation can be seen. For example, while the word “funny”

Table 3 Tag rankings for MovieLens dataset

Popularity	CPFSM
SCI-FI	
Matrix, The (1999)	Interstellar (2014)
Star Wars: Episode IV...	Matrix, The (1999)
Star Wars: Episode V...	Inception (2010)
Star Wars: Episode VI...	Arrival (2016)
Terminator 2...	Limitless (2011)
Jurassic Park (1993)	Alien (1979)
Back to the Future (1985)	Blade Runner 2049 (2017)
Twelve Monkeys...	Avengers, The (2012)
Alien (1979)	Edge of Tomorrow (2014)
Inception (2010)	V for Vendetta (2006)
ACTION	
Matrix, The (1999)	Matrix, The (1999)
Star Wars: Episode IV...	Terminator 2...
Fight Club (1999)	Star Wars: Episode V...
Star Wars: Episode V...	Star Wars: Episode IV...
Braveheart (1995)	Star Wars: Episode VI...
Raiders of the Lost Ark...	Raiders of the Lost Ark...
Star Wars: Episode VI...	Terminator, The (1984)
Terminator 2...	Saving Private Ryan (1998)
Jurassic Park (1993)	Die Hard (1988)
Saving Private Ryan (1998)	Blade Runner (1982)
ADVENTURE	
Star Wars: Episode IV...	Lord of the Rings I
Star Wars: Episode V...	Lord of the Rings II
Raiders of the Lost Ark...	Lord of the Rings III
Star Wars: Episode VI...	Star Wars: Episode IV...
Toy Story (1995)	Shrek (2001)
Lord of the Rings I	Pirates of the Caribbean...
Jurassic Park (1993)	Incredibles, The (2004)
Lord of the Rings III	WALL-E (2008)
Lord of the Rings II	Finding Nemo (2003)
Apollo 13 (1995)	Monsters, Inc. (2001)

Long titles are omitted due to space limitations

and “paranormal” are similar in the word embedding space, this edge decreases after the fit. On the other hand, “children” and “juvenile” do not have a strong similarity in terms of word vector, while the inference shows that this is a similar notion of preference.

7 Conclusion

We have proposed an approach to estimate the qualitative score from the binary features of products. Based on a natural assumption that an item with a better property is more

Table 4 Tag rankings for Goodbookss dataset

Tag counts	CPFSM
ADVENTURE	
Harry Potter	Oliver Twist
The Hunger Games	Treasure Island
Percy Jackson and the Olympians	Where the Red Fern Grows
The Lord of the Rings	Anne of Green Gables
Chronicles of Narnia	Black Beauty
Robert Langdon	The Adventures of Tom Sawyer
A Song of Ice and Fire	Heidi
A Series of Unfortunate Events	Little House
Dirk Pitt	The Swiss Family Robinson
Divergent	Robinson Crusoe
FANTASY	
Harry Potter	Chronicles of Narnia
Discworld	The Lord of the Rings
A Song of Ice and Fire	Charlotte's Web
Wheel of Time	The Hobbit
The Lord of the Rings	Harry Potter
Percy Jackson and the Olympians	A Christmas Carol
The Hobbit	Where the Wild Things Are
The Dresden Files	A Wrinkle in Time Quintet
His Dark Materials	The Book Thief
Twilight	Green Eggs and Ham
CHILDREN	
Harry Potter	Goodnight Moon
Chronicles of Narnia	Frog and Toad
A Series of Unfortunate Events	The Very Hungry Caterpillar Board Book
Little House	Number the Stars
Charlotte's Web	A Little Golden Book
Anne of Green Gables	Stone Soup
Where the Wild Things Are	Curious George
The Giving Tree	The Polar Express
A Wrinkle in Time Quintet	Madeline
Green Eggs and Ham	Where the Sidewalk Ends

popular among users who prefer that property, we have also proposed one discriminative and two generative models with which user preferences and item qualitative scores are inferred from user-item interactions. In these models, the space of the item qualitative score is constrained by the binary item features so that the score of each item and tag can only be nonzero when the item has the corresponding tag.

We have evaluated our models' performance by using two artificial and two real-world datasets. In the first experiment,

the performances of our model under the sparse transaction and noisy tag settings have been demonstrated by using the two artificial datasets. We have also evaluated our models' resolution for irrelevant tags using the MovieLens dataset and observed that they outperform the baseline model. Finally, we have discussed the tag rankings and tag correlations obtained from the two real-world datasets.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long

Fig. 6 Topic correlation graph for Goodbooks dataset. Gray lines represent edges with strong word correlations, while orange lines represent strong posterior correlations. Thick red lines represent if both word and posterior correlations are high

as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Geng X (2016) Label distribution learning. *IEEE Trans Knowl Data Eng* 28(7):1734–1748
- Shao R, Xu N, Geng X (2018) Multi-label learning with label enhancement. In: *ICDM*. IEEE, pp 437–446
- Goldberg D, Nichols DA, Oki BM, Terry DB (1992) Using collaborative filtering to weave an information tapestry. *Commun ACM* 35(12):61–70. <https://doi.org/10.1145/138859.138867>
- Resnick P, Iacovou N, Suchak M, Bergstrom P, Riedl J (1994) Grouplens: an open architecture for collaborative filtering of netnews. In: Smith JB, Smith FD, Malone TW (eds) *CSCW '94*, proceedings of the conference on computer supported cooperative work, Chapel Hill, NC, USA, October 22–26, 1994. ACM, pp 175–186. <http://doi.acm.org/10.1145/192844.192905>
- Shardanand U, Maes P (1995) Social information filtering: algorithms for automating “word of mouth”. In: Katz IR, Mack RL, Marks L, Rosson MB, Nielsen J (eds) *Human factors in computing systems*, CHI '95 conference proceedings, Denver, Colorado, USA, May 7–11, 1995. ACM/Addison-Wesley, pp 210–217. <http://doi.acm.org/10.1145/223904.223931>
- Sarwar B, Karypis G, Konstan J, Riedl J (2000) Application of dimensionality reduction in recommender system—a case study. Minnesota Univ Minneapolis Dept of Computer Science, Tech. Rep
- Kim D, Yum B (2005) Collaborative filtering based on iterative principal component analysis. *Expert Syst Appl* 28(4):823–830. <https://doi.org/10.1016/j.eswa.2004.12.037>
- Paterek A (2007) Improving regularized singular value decomposition for collaborative filtering. In: *Proceedings of KDD cup and workshop*, 2007, pp 5–8
- Bell RM, Koren Y, Volinsky C (2007) Modeling relationships at multiple scales to improve accuracy of large recommender systems. In: Berkhin P, Caruana R, Wu X (Eds) *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining*, San Jose, California, USA, August 12–15, 2007. ACM, pp 95–104. <http://doi.acm.org/10.1145/1281192.1281206>
- Salakhutdinov R, Mnih A, Hinton GE (2007) Restricted Boltzmann machines for collaborative filtering. In: Ghahramani Z (ed) *Machine learning, proceedings of the twenty-fourth international conference (ICML 2007)*, Corvallis, Oregon, USA, June 20–24, ser. ACM international conference proceeding series, 2007, vol 227. ACM, pp 791–798. <http://doi.acm.org/10.1145/1273496.1273596>
- Takács G, Pilászy I, Németh B, Tikk D (2007) Major components of the gravity recommendation system. *SIGKDD Explor* 9(2):80–83. <https://doi.org/10.1145/1345448.1345466>
- Koren Y (2008) Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: *KDD*, Li Y, Liu B, Sarawagi S (eds) ACM, pp 426–434. <http://doi.acm.org/10.1145/1401890.1401944>
- Pan R, Zhou Y, Cao B, Liu NN, Lukose RM, Scholz M, Yang Q (2008) One-class collaborative filtering. In: *ICDM*. IEEE

- Computer Society, pp 502–511. <https://doi.org/10.1109/ICDM.2008.16>
14. Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: ICDM. IEEE Computer Society, 2008, pp 263–272. <https://doi.org/10.1109/ICDM.2008.22>
 15. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. *J Am Soc Inf Sci* 41(6):391–407
 16. Hofmann T (1999) Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., pp 289–296
 17. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *J Mach Learn Res* 3(Jan):993–1022
 18. Chang J, Blei D (2009) Relational topic models for document networks. In: Artificial intelligence and statistics, pp 81–88
 19. Blei DM, Lafferty JD et al (2007) A correlated topic model of science. *Ann Appl Stat* 1(1):17–35
 20. Griffiths TL, Jordan MI, Tenenbaum JB, Blei DM (2004) Hierarchical topic models and the nested Chinese restaurant process. In: Advances in neural information processing systems, pp 17–24
 21. Li W, McCallum A (2006) Pachinko allocation: Dag-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on machine learning. ACM, pp 577–584
 22. Li C, Wang H, Zhang Z, Sun A, Ma Z (2016) Topic modeling for short texts with auxiliary word embeddings. In: Proceedings of the 39th international ACM SIGIR conference on research and development in information retrieval. ACM, pp 165–174
 23. Xun G, Li Y, Zhao WX, Gao J, Zhang A (2017) A correlated topic model using word embeddings. In: IJCAI, pp 4207–4213
 24. Yao L, Zhang Y, Wei B, Jin Z, Zhang R, Zhang Y, Chen Q (2017) Incorporating knowledge graph embeddings into topic modeling. In: Thirty-first AAAI conference on artificial intelligence
 25. Hu Z, Luo G, Sachan M, Xing EP, Nie Z (2016) Grounding topic models with knowledge bases. In: IJCAI, vol 16, pp 1578–1584
 26. Andrzejewski D, Zhu X, Craven M (2009) Incorporating domain knowledge into topic modeling via dirichlet forest priors. In: Proceedings of the 26th annual international conference on machine learning. ACM, pp 25–32
 27. Doshi-Velez F, Wallace BC, Adams R (2015) Graph-sparse lda: a topic model with structured sparsity. In: Twenty-Ninth AAAI conference on artificial intelligence
 28. Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. In KDD. ACM, pp 448–456
 29. Symeonidis P, Nanopoulos A, Manolopoulos Y (2008) Tag recommendations based on tensor dimensionality reduction. In: Proceedings of the 2008 ACM conference on recommender systems. ACM, pp 43–50
 30. Rendle S, Schmidt-Thieme L (2010) Pairwise interaction tensor factorization for personalized tag recommendation. In: Proceedings of the third ACM international conference on web search and data mining. ACM, pp 81–90
 31. He X, Zhang H, Kan M, Chua T (2016) Fast matrix factorization for online recommendation with implicit feedback. In: SIGIR, Perego R, Sebastiani F, Aslam JA, Ruthven I, Zobel J (eds) ACM, pp 549–558. <http://doi.acm.org/10.1145/2911451.2911489>
 32. Chen J, Zhu J, Wang Z, Zheng X, Zhang B (2013) Scalable inference for logistic-normal topic models. In: Advances in neural information processing systems, pp 2445–2453
 33. Hoyer PO (2004) Non-negative matrix factorization with sparseness constraints. *J Mach Learn Res* 5(Nov):1457–1469
 34. Donoho D, Stodden V (2004) When does non-negative matrix factorization give a correct decomposition into parts? In: Advances in neural information processing systems, pp 1141–1148
 35. Laurberg H, Christensen MG, Plumbley MD, Hansen LK, Jensen SH (2008) Theorems on positive data: on the uniqueness of NMF. *Computational intelligence and neuroscience* 2008
 36. Huang K, Sidiropoulos ND, Swami A (2013) Non-negative matrix factorization revisited: uniqueness and algorithm for symmetric decomposition. *IEEE Trans Signal Process* 62(1):211–224
 37. Vavasis SA (2009) On the complexity of nonnegative matrix factorization. *SIAM J Optim* 20(3):1364–1377
 38. Bergstra J, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira FCN, Weinberger KQ (eds) Advances in neural information processing systems 24: 25th annual conference on neural information processing systems. Proceedings of a meeting held 12–14 December 2011, Granada, Spain, 2011, pp 2546–2554. <http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization>
 39. Akiba T, Sano S, Yanase T, Ohta T, Koyama M (2019) Optuna: a next-generation hyperparameter optimization framework. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G (eds) Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019. ACM, pp 2623–2631. <https://doi.org/10.1145/3292500.3330701>
 40. Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) BPR: bayesian personalized ranking from implicit feedback. In: Bilmes JA, Ng AY (eds) UAI 2009, Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, Montreal, QC, Canada, June 18–21. AUAI Press, 2009, pp 452–461. https://dslpipt.org/uai/displayArticleDetails.jsp?mmnu=1&smnu=2&article_id=1630&proceeding_id=25
 41. Vig J, Sen S, Riedl J (2012) The tag genome: encoding community knowledge to support novel interaction. *ACM Trans Interactive Intell Syst (TiiS)* 2(3):13
 42. Zajac Z (2017) Goodbooks-10k: a new dataset for book recommendations, FastML
 43. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. *Trans Assoc Comput Linguistics* 5:135–146