A Joint Optimization Approach for Personalized Recommendation Diversification

Xiaojie Wang[1], Jianzhong Qi[2], Kotagiri Ramamohanarao[2], Yu Sun[3], Bo Li[4], and Rui Zhang[2]* .

[1,2]The University of Melbourne, [3]Twitter Inc., [4]University of Illinois at Urbana Champaign.
[1]xiaojiew1@student.unimelb.edu.au, [2]{jianzhong.qi,rui.zhang,kotagiri}@unimelb.edu.au,
[3]ysun@twitter.com, [4]lxbosky@gmail.com.

Abstract. In recommendation systems, items of interest are often classified into categories such as genres of movies. Existing research has shown that diversified recommendations can improve real user experience. However, most existing methods do not consider the fact that users' levels of interest (i.e., user preferences) in different categories usually vary, and such user preferences are not reflected in the diversified recommendations. We propose an algorithm that considers user preferences for different categories when recommending diversified results, and refer to this problem as personalized recommendation diversification. In the proposed algorithm, a model that captures user preferences for different categories is optimized jointly toward both relevance and diversity. To provide the proposed algorithm with informative training labels and effectively evaluate recommendation diversity, we also propose a new personalized diversity measure. The proposed measure overcomes limitations of existing measures in evaluating recommendation diversity: existing measures either cannot effectively handle user preferences for different categories, or cannot evaluate both relevance and diversity at the same time. Experiments using two real-world datasets confirm the superiority of the proposed algorithm, and show the effectiveness of the proposed measure in capturing user preferences.

1 Introduction

In most recommendation systems, items are classified by predefined categories, e.g., genres of movies or styles of musics. Recent studies show that users' interests often spread into several genres [20, 21] (for ease of presentation, we will simply use genres to represent categories in the following). However, many existing algorithms (e.g., [7, 8]) only try to optimize toward recommendation accuracy or item relevance, which is not optimal to cover users' diverse interests. In fact, the objectives of relevance and diversity are largely orthogonal, i.e., optimizing toward relevance may recommend very similar items, while optimizing toward diversity may present less relevant items. Recommendation diversification algorithms aim to achieve these two objectives at the same time and recommend diverse items

[*] Corresponding author

Table 1. Three lists of recommended movies in movie recommendations.

	Recommendations by three different ranking measures		
Rank	Non-diverse recomm.	Diverse without user pref.	Diverse with user pref.
1	First Shot (★)	First Shot (★)	First Shot (★)
2	Rapid Fire (★)	Snow Angels (⊛)	Snow Angels (⊛)
3	Black Dawn (★)	Rapid Fire (★)	Rapid Fire (★)
4	Shadow Man (★)	iss Potter (⊛)	Black Dawn (★)
Count	#Action=4 #Drama=0	#Action=2 #Drama=2	#Action=3 #Drama=1

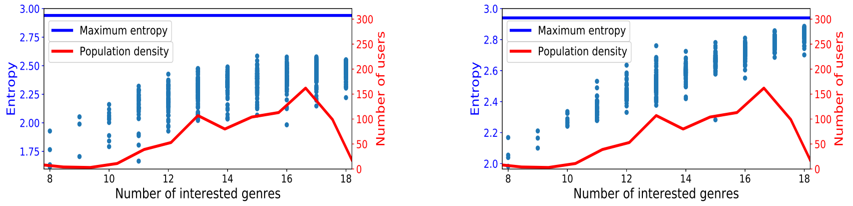
[1] Star (★) stands for action movies and asterisk (⊛) stands for drama movies.

with high relevance. Existing work in this area either separates relevance and diversity optimization [18], or does not explicitly consider the personalization in genre preferences [3, 5, 19] as discussed below.

Users usually have varied preferences over different genres [18]. High variances in such genre preferences require highly personalized recommendation diversification algorithms, which aim to present diverse recommendations catering to individual user's genre preference [18]. For example, Table 1 shows three lists of movies recommended to a user interested in both action and drama movies. The movies under the "non-diverse recomm." column are all action movies, which are not diverse in terms of genres. The movies under the "diverse without user pref." and "diverse with user pref." columns resolve this issue by also presenting drama movies. Suppose that the user prefers action movies. The "diverse without user pref." column treats the two genres equally (recommending two action movies and two drama movies) and does not consider the user's genre preference. The "diverse with user pref." column in this case presents a better recommendation, i.e., personalized diverse recommendations, which is the aim of this paper.

Toward this end, we propose a *personalized diversification algorithm* to jointly optimize both relevance and diversity and explicitly consider personalized genre preferences in diversification. The proposed algorithm iteratively selects the item that maximizes a function (i.e. ranking function) of two components: one models a user's rating for an item and the other models the user's genre preference for the item. The two components are collaborated by a joint optimization method to recommend items as accurately as possible (accurate rating prediction) and make an item list as personalized diverse as possible (personalized diverse ranking). The joint optimization method enables the personalized diversification algorithm to use the true ratings and pre-determined item rankings as sources of training information, where the item rankings indicate which item should be selected for personalized diverse recommendations given a selected item list.

To provide effective item rankings (i.e., training labels) to our algorithm, we need to measure the diversity of recommendations for each user, i.e., *personalized diversity*. Existing measures have limitations in evaluating personalized diversity: they either cannot handle the genre preferences of a user [4], or ignore the minor interests of a user [1], or cannot evaluate both relevance and diversity at the



(a) The frequency-based user preference. (b) The rating-based user preference.

Fig. 1. User preference analysis on a movie rating dataset (#genres=18).

same time [18]. To overcome these limitations, we propose a new *personalized diversity measure*, which evaluates an item list based on user preferences for the covered genres of the list. This makes the item list having the highest score under our measure (i.e., the ideal list) has a desired property [18]: each genre is represented according to personalized genre preferences in the list.

The main contributions of this paper include: (1) We propose a novel recommendation diversification algorithm which can learn a ranking function by jointly optimizing the relevance and diversity. (2) We also propose a personalized diversity measure that can effectively evaluate personalized diversity of recommendations. (3) Experiments using real-world datasets of different domains show that the proposed algorithm outperforms several baseline methods and the proposed measure is more effective in capturing personalized genre preferences.

2 Problem Formulation

We assume that items to be recommended are categorized into genres. Let $\mathcal{X} = \{x_n\}_{n=1}^{N}$ be an item set, $\mathcal{G} = \{g_k\}_{k=1}^{K}$ be a genre set, $\mathbf{R} \in \mathbb{R}^{U \times N}$ be a rating matrix ($\mathbf{R}_{u,n}$ is the rating of user u for item x_n), $\mathbf{J} \in \mathbb{R}^{N \times K}$ be the genre information for items $\mathcal{X}$ ($\mathbf{J}_{n,g} = 1$ if item x_n is with genre g and $\mathbf{J}_{n,g} = 0$ otherwise). We define the problem of personalized recommendation diversification as:

Definition 1 (Personalized Recommendation Diversification). Given U users, N items, K genres, the rating matrix $\mathbf{R}$, the genre information $\mathbf{J}$, and a personalized diversity measure $\mathcal{M}$, the task is to generate the item list $\mathcal{Y}^u = [x_{y_1}, ..., x_{y_N}]$ that maximizes the measure $\mathcal{M}$ for each user u.

Intuitively, the problem is to consider **personalized genre preferences** (referred to as **user preferences** in the following for brevity) in diversification.

We consider two formulations of modeling user preferences. Let $\mathcal{X}^u$ be the item set rated by user u and $\mathcal{X}_g^u \subseteq \mathcal{X}^u$ be the subset of items with genre g. The *frequency-based user preference* is given by $p_g^u \propto |\mathcal{X}_g^u|/|\mathcal{X}^u|$ ($g \in \mathcal{G}$) [18]. Here, p_g^u is the user preference for genre g, which is proportional to the percentage of rated items with genre g. To consider the scale of ratings, we define the *rating-based user preference* as $q_g^u \propto \sum_{n:x_n \in \mathcal{X}_g^u} \mathbf{R}_{u,n} / \sum_{n:x_n \in \mathcal{X}^u} \mathbf{R}_{u,n}$ ($g \in \mathcal{G}$). Here, q_g^u is proportional to the sum of ratings for items with genre g, $\sum_{n:x_n \in \mathcal{X}_g^u} \mathbf{R}_{u,n}$, over the sum of ratings for items with any genre, $\sum_{n:x_n \in \mathcal{X}^u} \mathbf{R}_{u,n}$.

We identify a few key characteristics of user preferences using a movie rating dataset detailed in Section 5. We show the entropy of user preferences against the number of interested genres (those genres covered by $\mathcal{X}^u$) in Figure 1, where each user is a blue dot. For both user preference formulations, we find that: (1) Users prefer different degrees of diversity: the number of interested genres varies from 8 to 18 and the entropy of user preferences varies from 1.5 to 3.0 across users; (2) Users have varied preferences for different genres: no single user reaches the maximum entropy line where all genres are of the same interest to a user. These findings motivate us to consider user preferences in diversification.

3 Personalized Diversification Algorithms

In theory, optimizing a diversity measure is NP-hard [1], and a greedy strategy is often adopted [3]: at iteration r, $r-1$ items $\mathcal{Y}_{r-1}$ have been selected. A marginal score function $s(x_n, \mathcal{Y}_{r-1})$ is used to select the next best item, which is then added to $\mathcal{Y}_{r-1}$. Two methods for modeling $s(x_n, \mathcal{Y}_{r-1})$ are presented as follows.

3.1 Personalized Diversification Algorithm by Greedy Re-ranking

A naive method is to use a re-ranking strategy which greedily selects next items based on predicted ratings, which is called *personalized diversification algorithm based on greedy re-ranking* (PDA-GR). It consists of: (1) A prediction phrase uses matrix factorization to predict ratings $\{\hat{\mathbf{R}}_{u,n}\}_{x_n \in \mathcal{X}}$; (2) A re-ranking phrase uses a training set to estimate user preferences $\{\hat{p}_g^u\}_{g \in \mathcal{G}}$, and a heuristic-based marginal score function to re-rank. Using the genre information $\mathbf{J}$, the marginal score function is defined as a combination of a rating component $f(\hat{\mathbf{R}}_{u,n})$, which models a user's rating for item x_n, and a genre preference component $\mathbf{J}_{n,g}(\hat{p}_g^u)^{C_g(r-1)}$, which models the user's genre preference of item x_n:

$$s(x_n, \mathcal{Y}_{r-1}) = \sum_{g \in \mathcal{G}} f(\hat{\mathbf{R}}_{u,n}) \cdot \mathbf{J}_{n,g}(\hat{p}_g^u)^{C_g(r-1)} \tag{1}$$

Here, $f(r) = 2^r$, and $C_g(r-1)$ is the number of previous items with genre g. PDA-GR is sub-optimal because it divides optimizing accurate rating prediction and personalized diverse ranking into two separate phrases.

3.2 Personalized Diversification Algorithm by Joint Optimization

To tackle the sub-optimality, we propose a *personalized diversification algorithm based on joint optimization* (PDA-JO), which can optimize both accurate rating prediction and personalized diverse ranking simultaneously.

For user u, let $\mathbf{p}_u \in \mathbb{R}^F$ be the embedding and $\mathbf{b}_u \in \mathbb{R}$ be the bias. For item x_n, let $\mathbf{q}_n \in \mathbb{R}^F$ be the embedding and $\mathbf{b}_n \in \mathbb{R}$ be the bias. The rating for item x_n is predicted by $\hat{\mathbf{R}}_{u,n} = \mathbf{p}_u^\intercal \mathbf{q}_n + \mathbf{b}_u + \mathbf{b}_n$. We use a parameter μ to alleviate the error of rating prediction. The marginal score function is defined as:

$$s(x_n, \mathcal{Y}_{r-1}) = \sum_{g \in \mathcal{G}} f(\hat{\mathbf{R}}_{u,n} + \mu) \cdot \mathbf{J}_{n,g}(\hat{p}_g^u)^{C_g(r-1)} \tag{2}$$

```
Algorithm 1: Personalized Diversification Algorithm by Joint Optimization
Input: users $\mathcal{U}$, items $\mathcal{X}$, ratings $\mathbf{R}$, a personalized diversity measure $\mathcal{M}$
Pre-train $\{\mathbf{p}_u, \mathbf{b}_u\}_{u \in \mathcal{U}}, \{\mathbf{q}_n, \mathbf{b}_n\}_{x_n \in \mathcal{X}}$ based on $\mathbf{R}$
while PDA-JO has not converge do
    $\mathcal{Z} \leftarrow \varnothing, \mathcal{B} \leftarrow \varnothing$          ◇ $\mathcal{Z}$ is sampled item lists and $\mathcal{B}$ is training instances
    for each user $u$ in $\mathcal{U}$ do
        for length $l$ from $0$ to $|\mathcal{X}| - 1$ do
            Add the ideal list of length $l$ under the measure $\mathcal{M}$ into $\mathcal{Z}$
            Sample $S$ non-ideal lists of length $l$ and add them into $\mathcal{Z}$
    for item list $\mathcal{Y}$ in $\mathcal{Z}$ do
        for item pair $(x_m, x_n)$ from $\mathcal{X} \backslash \mathcal{Y}$ do
            if $M(\mathcal{Y} + [x_m]) > M(\mathcal{Y} + [x_n])$ then $L \leftarrow 1$
            else $L \leftarrow 0$
            Add $(\mathcal{Y}, x_m, x_n, y = (\mathbf{R}_{u,m}, \mathbf{R}_{u,n}, L))$ into $\mathcal{B}$
    for mini-batch $b$ in $\mathcal{B}$ do
        Update $\{\mathbf{p}_u, \mathbf{b}_u\}_{u \in \mathcal{U}}, \{\mathbf{q}_n, \mathbf{b}_n\}_{x_n \in \mathcal{X}}, \mu$ based on Equation 3
```

Here, $f(r) = 2^r$, $\mathbf{J}_{n,g} = 1$ if item x_n is with genre g and $\mathbf{J}_{n,g} = 0$ otherwise, p_g^u is the preference of user u for genre g, and $C_g(r-1)$ is the number of previous items with genre g. $\{\mathbf{p}_u, \mathbf{b}_u\}_{u \in \mathcal{U}}$, $\{\mathbf{q}_n, \mathbf{b}_n\}_{x_n \in \mathcal{X}}$, and μ are learnable parameters.

We define a training instance for user u as $(\mathcal{Y}, x_m, x_n, y)$ where $\mathcal{Y}$ is selected items, x_m and x_n are two candidate items, and $y = (\mathbf{R}_{u,m}, \mathbf{R}_{u,n}, L)$. Here, $\mathbf{R}_{u,m}$ and $\mathbf{R}_{u,n}$ are the true ratings for items x_m and x_n. Training label L indicates which item ranking is better under the measure $\mathcal{M}$: $L = 1$ if $\mathcal{M}(\mathcal{Y} + [x_m]) > \mathcal{M}(\mathcal{Y} + [x_n])$ and $L = 0$ otherwise. The probability of $L = 1$ is $P = \sigma(s(x_m, \mathcal{Y}) - s(x_n, \mathcal{Y}))$, where $\sigma(\cdot)$ is the sigmoid function. The loss function of our algorithm consists of a relevance loss $\mathcal{L}_r$ a personalized diversity loss $\mathcal{L}_d$:

$$\mathcal{L} = \underbrace{0.5[(\hat{\mathbf{R}}_{u,m} - \mathbf{R}_{u,m})^2 + (\hat{\mathbf{R}}_{u,n} - \mathbf{R}_{u,n})^2]}_{\text{The relevance loss: } \mathcal{L}_r} - \underbrace{D[L \log P + (1-L)\log(1-P)]}_{\text{The personalized diversity loss: } \mathcal{L}_d}$$

Here, D balances between accurate rating prediction (loss $\mathcal{L}_r$) and personalized diverse ranking (loss $\mathcal{L}_d$). We use L_2 regularization to regularize the model.

The model is trained by stochastic gradient descent with gradient given by:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{p}_u} &= (e_{u,m}\mathbf{q}_m - e_{u,n}\mathbf{q}_n) + DE\{\sum_{g \in \mathcal{G}} d_g f'(\hat{\mathbf{R}}_{u,m} + \mu)\mathbf{q}_m - \sum_{g \in \mathcal{G}} d_g f'(\hat{\mathbf{R}}_{u,n} + \mu)\mathbf{q}_n\} \\
\frac{\partial \mathcal{L}}{\partial \mathbf{q}_l} &= e_{u,l}\mathbf{p}_u + DE\{\sum_{g \in \mathcal{G}} d_g f'(\hat{\mathbf{R}}_{u,l} + \mu)\mathbf{p}_u\} \ l \in \{m, n\} \\
\frac{\partial \mathcal{L}}{\partial \mathbf{b}_u} &= (e_{u,m} - e_{u,n}) + DE\{\sum_{g \in \mathcal{G}} d_g f'(\hat{\mathbf{R}}_{u,m} + \mu) - \sum_{g \in \mathcal{G}} d_g f'(\hat{\mathbf{R}}_{u,n} + \mu)\} \\
\frac{\partial \mathcal{L}}{\partial \mathbf{b}_l} &= e_{u,l} + DE\{\sum_{g \in \mathcal{G}} d_g f'(\hat{\mathbf{R}}_{u,l} + \mu)\} \ l \in \{m, n\} \\
\frac{\partial \mathcal{L}}{\partial \mu} &= DE\{\sum_{g \in \mathcal{G}} d_g f'(\hat{\mathbf{R}}_{u,m} + \mu) - \sum_{g \in \mathcal{G}} d_g f'(\hat{\mathbf{R}}_{u,n} + \mu)\}
\end{aligned} \tag{3}$$

Algorithm 2: Building the ideal list for p-nDCG

```
Input: user $u$, items $\mathcal{X} = \{x_n\}_{n=1}^{N}$, user ratings $\mathbf{R}$, genre information $\mathbf{J}$
Output: ideal list $\mathcal{Y} = \{x_{y_n}\}_{n=1}^{N}$
Estimate the user preferences $\{p_g^u\}_{g \in \mathcal{G}}$ based on the genre information $\mathbf{J}$
$\mathcal{Y}_0 \leftarrow \varnothing$                    ◇ a selected item list
for $r = 1, ..., N$ do
    $x_m \leftarrow \arg\max_{x_n \in \mathcal{X} \setminus \mathcal{Y}_{r-1}}(\text{p-nDCG}(\mathcal{Y}_{r-1} + [x_n]) - \text{p-nDCG}(\mathcal{Y}_{r-1}))$
    $\mathcal{Y}_r \leftarrow \mathcal{Y}_{r-1} + [x_m]$
$\mathcal{Y}^u \leftarrow \mathcal{Y}_N$
```

Here, $e_{u,l} = \hat{\mathbf{R}}_{u,l} - \mathbf{R}_{u,l}$ $l \in \{m, n\}$, $E = P - L$, and $d_g = (\hat{p}_g)^{C_g(r-1)}$. The total number of training instances is $\Theta(MN!)$. To speed up training, we use a sampling method similar to the negative sampling [9]: both ideal lists and a number of sampled non-ideal lists under measure $\mathcal{M}$ are used to estimate the gradient. The overall procedure of the joint optimization method is summarized in Algorithm 1. The model is first pre-trained by training ratings. Then, we sample S non-ideal lists with a certain length $l \in [0, N-1]$ for each user and update the parameters with the gradient given by Equation 3.

Time Complexity. The training time complexity is $\Theta(E \cdot M \cdot S \cdot N^2 \cdot T)$, where E is the number of epoches, S is the number of sampled non-ideal item lists. $T = \max\{F, K\}$ is the time complexity of computing the marginal score function. The test time complexity is $\Theta(N^2 \cdot T)$ for each user.

4 Personalized Diversity Measure

Existing measures have limitations in evaluating personalized recommendation diversity. Therefore, we proposed a personalized diversity measure in this section.

4.1 Limitations of Existing Diversity Measures

Our goals are to recommend items that (I) cover a user's interested genres, and (II) have a genre distribution satisfying the user's preference for different genres. Existing diversity measures cannot serve our goals: (1) α-nDCG [4] does not model user preferences (or intent probabilities). (2) IA measures [1] tend to favor the major interests and ignore the minor interests of a user [12]. (3) One of the goals of D♯-measures [12] is to recommend items that cover as many genres (or intents) as possible, but not to optimize toward individual user's preference.

4.2 Formulation of Personalized Diversity Measure

To overcome these limitations, we propose a personalized diversity measure. Our measure is motivated by α-nDCG [4], which discounts the gain of redundant items by a constant $\alpha \in [0, 1]$. Users often have varied preferences for different

Table 2. The movielens 100k dataset (ML-100K) and the million song dataset (MSD)

Stat. Data	#users	#items	#ratings	#genres	Range	Sparsity
ML-100K	943	1,682	100,000	18	1-5	6.30%
MSD	1,217	2,051	88,078	15	1-225	3.53%

genres. A constant cannot model such variances. Intuitively, redundancy under more preferred genres is better than redundancy under less preferred genres.

Let $J_g(r) = 1$ if the item at rank r is labeled with genre g and $J_g(r) = 0$ otherwise, and $C_g(r) = \sum_{k=1}^{r} J_g(k)$. Based on the preference of user u $\{p_g^u\}_{g \in \mathcal{G}}$, we define the *personalized novelty-biased gain* (PNG) for the item at rank r as:

$$PNG(r) = \sum_{g \in \{g\}} h(r) \cdot J_g(r) (p_g^u)^{C_g(r-1)} \tag{4}$$

Here, $h(r) = (2^r - 1)/2^{r_{max}}$. PNG models the marginal gain of an item after a user has seen previous items. We define *p-nDCG* at cutoff C as:

$$p\text{-}nDCG@C = \frac{\sum_{r=1}^{C} PNG(r)/\log(r+1)}{\sum_{r=1}^{C} PNG^*(r)/\log(r+1)} \tag{5}$$

Here, PNG^* is PNG of the ideal list built by Algorithm 2. The algorithm iteratively selects the item that maximizes the p-nDCG score of current item list based on the true ratings and user preferences.

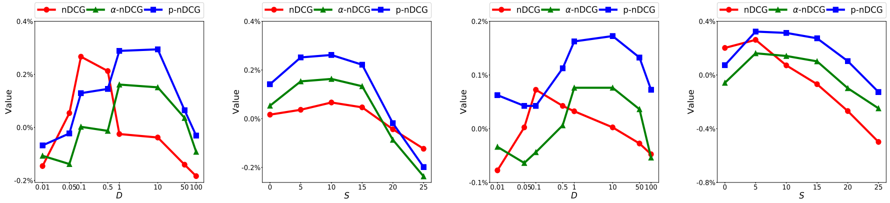
Theoretical Analysis. p-nDCG is effective in capturing user preferences: item lists with a high p-nDCG score tend to contain more items with more preferred genres. To see this, we analyze the ideal list under p-nDCG. If genre g is under-represented in the list, i.e. p_g is high while C_g is low, the PNG for a relevant item with genre g will be large. This makes p-nDCG select more relevant items with genre g as next items. The selection process reaches an equilibrium when each genre is represented according to user preferences:

$$(p_{g_1})^{C_{g_1}} \equiv (p_{g_2})^{C_{g_2}} \ (g_1, g_2 \in \mathcal{G}) \quad \Rightarrow \quad C_g \propto \log(p_g) \ (g \in \mathcal{G}) \tag{6}$$

The ideal list is effective in reflecting user preferences: the number of items with genre g (C_g) is positively correlated with the preference for genre g (p_g) in the list. This is a desired property for personalized recommendation diversification [18]: each genre needs to be represented according to user preferences in an item list.

5 Experiments

We experiment with the movielens 100k dataset (ML-100K) [6] and the million song dataset (MSD) [2]. ML-100K is a movie rating dataset. It contains 100,000 ratings on 1,682 movies from 943 users. MSD contains music play counts. We use a subset of MSD containing the playing counts of the songs associated properly to one of the predefined genres. This subset contains 88,078 playing counts on 2,051 songs from 1,217 users. The two datasets are summarized in Table 2.



(a) Varied D on ML-100K. (b) Varied S on ML-100K. (c) Varied D on MSD. (d) Varied S on MSD.

Fig. 2. Performances of PDA-JO with varied parameters after z-normalization.

We try both formulations of user preferences and obtain similar results in the experiments. We only show the results using the frequency-based user preference due to the page limit. We use normalized discounted cumulative gain (nDCG), α-nDCG ($\alpha = 0.5$) [4], and the proposed p-nDCG to evaluate algorithm performances. All these measures are computed at cutoff $C = 10$.

5.1 Experiments on Algorithms

The compared methods include **MF** [7], **MMR** [3], **PM-2** [5], and **LTR-N** [19]. We use 5-fold cross validation to tune parameters for all algorithms.

Effects of Parameters. In Figure 2, we present the effects of tuning (1) D varied from 0.01 to 100 , and (2) S varied from 0 to 25. We apply the z-normalization method to amplify the effects. The proposed PDA-JO performs best when $(D, S) = (1, 10)$ on ML-100K and $(D, S) = (10, 5)$ on MSD. Figures 2(a) and 2(c) show the effects of D on ML-100K ($S = 10$) and MSD ($S = 15$). The performance of PDA-JO increases with the growth of D ($0.1 \leqslant D \leqslant 10$), after which the performance decreases under α-nDCG and p-nDCG. This is because: (1) If D is small, PDA-JO is biased toward rating prediction and disregard diverse ranking, which will degrade the performance under diverse measures. (2) If D is large, rating prediction is less accurate, which will in turn degrade the performance because diverse ranking relies on rating prediction (see Equation 2). The influence of D is stable when $1 \leqslant D \leqslant 10$. Figures 2(b) and 2(d) show the effects of S on ML-100K ($D = 1$) and MSD ($D = 10$). The proposed PDA-JO performs better as S increases ($0 \leqslant S \leqslant 10$), but a performance decrease occurs when $S \geqslant 15$. The overall difference when varying D and S is less than 0.8%, which indicates that PDA-JO is a robust framework.

Comparison of Algorithms. Table 3 compares the performances of all algorithms on ML-100K and MSD. The proposed PDA-JO performs best on both datasets under all three measures. The improvement of PDA-JO over baseline methods is significant based on two-tailed paired t-test. We compare all methods in the following aspects: (1) Personalized diversification methods (PDA-GR and PDA-JO) outperform non-personalized diversification methods (MMR, PM-2, and LTR-N) on all three measures. (2) Heuristic-based methods (MMR and PM-2) sacrifice relevance to boost diversity, while learning-based methods

Table 3. Performance comparison of algorithms on ML-100K and MSD. For MMR and PM-2, the subscript is the parameter achieving the best score on validation set.

	Performance on MK-100K			Performance on MSD		
Method	nDCG	α-nDCG	p-nDCG	nDCG	α-nDCG	p-nDCG
MF	0.7206	0.6035	0.5799	0.6061	0.4728	0.5001
$\text{MMR}_{0.7}$	0.6944	0.6206 (2.82%)	0.6172 (6.44%)	0.6081	0.4803 (1.58%)	0.5068 (1.33%)
$\text{PM-2}_{0.5}$	0.6829	0.6759 (11.98%)	0.6525 (12.53%)	0.5895	0.4954 (4.77%)	0.5179 (3.54%)
LTR-N	0.7301	0.7134 (18.21%)	0.7017 (21.00%)	0.6230	0.4997 (5.70%)	0.5246 (4.89%)
PDA-GR	0.7283	0.7782 (28.93%)	0.7690 (32.61%)	0.6295	0.5430 (14.85%)	0.5665 (13.27%)
PDA-JO	**0.7417**	**0.7846** (29.99%)	**0.7778** (34.13%)	**0.6309**	**0.5579** (18.00%)	**0.5808** (16.14%)

(LTR-N and PDA-JO) can improve both relevance and diversity. (3) PDA-JO is consistently better than PDA-GR for all measures on both datasets.

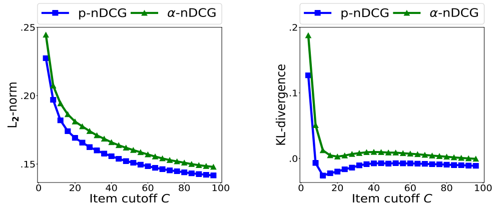
5.2 Experiments on Measures

We compare the ideal lists of p-nDCG and α-nDCG ($\alpha = 0.5$) on ML-100K as follows: (1) For each user, we randomly split ratings into a training set (80%) and a test set (20%). We also use time-based split (the most current 20% are used for testing), and the results are similar; (2) We use the training set to build the ideal list of p-nDCG (α-nDCG) by Algorithm 2. Here, the user preferences used to compute the p-nDCG score are obtained using the training set.

Satisfying User Preferences. We show that the ideal list of p-nDCG is more effective than α-nDCG in reflecting user preferences. We compute genre distribution P_p (P_α) of p-nDCG (α-nDCG) by applying user preference formulations to the top-C ranked items in the ideal list. The ground-truth user preference P^* is obtained using the test set. We compute the distance between P^* and P_p (P_α) using KL-divergence or L_2-norm, and average all distances across users. We plot the average distance against item cutoff in Figure 3. We find that compared with α-nDCG, the genre distribution of the top-C ranked items by p-nDCG consistently better satisfy user preferences, especially when cutoff C is small.

Rank Correlation. We use Kendall's τ to measure rank correlation between the ideal lists of p-nDCG and α-nDCG. The results of averaging Kendall's τ over the users who are interested in the same number of genres are shown in Figure 4. We find that as the number of interested genres increases, the rank correlation decreases. This is because when a user's interested genres are of the same interest to the user, p-nDCG reduces to α-nDCG. As the number of interested genres grows, the probability that a user has the same preference for different genres decreases. This causes p-nDCG and α-nDCG to produce less similar item lists.

Case Study. We use a real user on ML-100K to illustrate the advantage of p-nDCG in Table 4. The ground-truth column (user preferences) is computed by applying the frequency-based user preference to the test set. We find that: (1) In terms of genre ranking, p-nDCG is more consistent (Kendall's $\tau = 0.89$) with the user preferences than α-nDCG (Kendall's $\tau = 0.39$); (2) The genre distribution of recommended items by p-nDCG is closer (L_2-norm $= 0.20$ using



(a) The L_2-norm. (b) The KL-divergence.

Fig. 3. The distance between the ground-truth user preferences and the genre distribution of the top-C ranked items by p-nDCG (α-nDCG), where the item cutoff $C \in [5, 100]$.

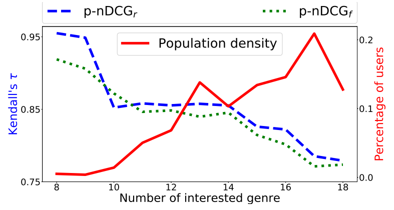


Fig. 4. Rank correlation (Kendall's τ) with α-nDCG, where p-nDCG$_f$ uses the frequency-based user preference while p-nDCG$_r$ uses the rating-based user preference.

the frequency-based user preference) to the user preferences than α-nDCG (L_2-norm $= 0.29$ using the frequency-based user preference).

6 Related Work

Before receiving attentions in recommendation systems (RS), the problem of diversity is studied in information retrieval (IR) [1, 3–5, 12, 19]. One difference between IR work and our work is that there is ground-truth for test item genres in our work (e.g., ML-100K provides the genre information of movies), but there is no such ground-truth for test document intents (analogous to item genres) in IR work. We explicitly incorporate such genre information into the diverse ranking model, which makes even the naive method effective. Another difference is that the embedding is trainable in our work, but the embedding is not trainable in IR work (it is pre-computed and fixed as relevance features) [19].

Diversity Measures. Several diversity measures are proposed in IR to evaluate the diversity [1, 4, 12]. They are not designed to evaluate the personalized diversity as discussed in Section 4.1. In RS, Smyth and McClave [13] define the dissimilarity-based diversity, i.e., the average dissimilarity between all pairs of the recommended items. Vargas et al. argue that the dissimilarity-based diversity is less likely to be perceived as diverse by users than the genre diversity [18]. They propose a Binomial framework to evaluate the genre diversity. The Binomial framework cannot evaluate the relevance (random recommendations may achieve high scores under this framework) and does not model the position of relevant item in an item list. It differs from our measure which evaluates both relevance and diversity and models the relevant item position.

Related Algorithms. Diversification algorithms can be categorized into heuristic-based and learning-based. Heuristic-based methods use some heuristic rules to re-rank the candidate items [3, 5, 21]. For example, Ziegler et al. propose to select the next item by linearly combining the relevance and the dissimilarity

Table 4. The ideal list of p-nDCG (α-nDCG) for a real user on ML-100K.

Pos. / Pref.	Top-10 items by $\boldsymbol{\alpha}$-nDCG											Ground-truth		Frequency-based		Rating-based	
	1	2	3	4	5	6	7	8	9	10	Ct.	Pref.	Rank	Pref.	Rank	Pref.	Rank
Comedy				◆	◆			◆			3	0.3789	1	0.1875	2	0.2000	2
Horror				◆		◆	◆				3	0.2756	2	0.1875	2	0.2000	2
Romance				◆	◆			◆			3	0.2067	3	0.1875	2	0.2000	2
Animation	◆					◆			◆	◆	4	0.0689	4	0.2500	1	0.2154	1
Adventure		◆					◆				2	0.0689	4	0.1250	5	0.1231	5
Thriller			◆								1	0.0011	6	0.0625	6	0.0615	6
Statistics														L$_2$ (**0.29**)	τ (**0.39**)	L$_2$ (**0.26**)	τ (**0.39**)

Pos. / Pref.	Top-10 items by **p-nDCG**											Ground-truth		Frequency-based		Rating-based	
	1	2	3	4	5	6	7	8	9	10	Ct.	Pref.	Rank	Pref.	Rank	Pref.	Rank
Comedy	◆					◆		◆	◆		4	0.3789	1	0.2353	1	0.2500	1
Horror	◆	◆						◆		◆	4	0.2756	2	0.2353	1	0.2500	1
Romance	◆					◆			◆		3	0.2067	3	0.1765	3	0.1842	3
Animation					◆		◆			◆	3	0.0689	4	0.1765	3	0.1447	4
Adventure		◆		◆							2	0.0689	4	0.1176	5	0.1184	5
Thriller			◆								1	0.0011	6	0.0588	6	0.0526	6
Statistics														L$_2$ (**0.20**)	τ (**0.89**)	L$_2$ (**0.17**)	τ (**0.93**)

[1] Diamond ◆ indicates the movie at a certain position is categorized as a certain genre.

[2] L$_2$ stands for the L$_2$-norm and τ stands for the Kendall's τ.

to the selected items based on an intra-list similarity measure [21]. Learning-based methods aim to learn a diverse ranking model from a training set [19]. For example, Xia et al. propose to learn a diverse ranking model by using neural networks to model the marginal novelty of candidate items.

The proposed algorithm is related to model-based collaborative filtering methods, which explain user ratings by factoring the ratings into user embedding and item embedding [7, 11]. Our algorithm borrows ideas from learning-to-rank methods [10], which overcome the problems with heuristic predefined ranking function. For example, Tran et. al propose to integrate deep neural networks into the learning-to-rank model [17]. Our algorithm is also related to intent tracking algorithms [14–16] in designing highly personalized recommendation systems: we aim to personalize at genre level while intent tracking algorithms personalize at intent level. However, none of these algorithms explicitly consider personalized genre preferences, which is the topic of our work.

7 Conclusion

We studied the problem of personalized recommendation diversification. A personalized diversification algorithm was proposed to incorporate user preferences and jointly optimize both relevance and diversity. To overcome limitations of existing measures, we proposed a personalized diversity measure to evaluate the personalized diversity of recommendations. Experiments using real-world datasets showed that the proposed algorithm outperforms baseline algorithms, including a state-of-the-art leaning-to-rank algorithm. The experiments also validated the effectiveness of the proposed measure in capturing user preferences.

8 Acknowledgment

This work is supported by Australian Research Council (ARC) Future Fellowships Project FT120100832 and Discovery Project DP180102050.

References

1. Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S.: Diversifying search results. In: WSDM. pp. 5–14. ACM (2009)
2. Bertin-Mahieux, T., Ellis, D.P., Whitman, B., Lamere, P.: The million song dataset. In: ISMIR. vol. 2, p. 10 (2011)
3. Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. pp. 335–336 (1998)
4. Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I.: Novelty and diversity in information retrieval evaluation. In: SIGIR. pp. 659–666. ACM (2008)
5. Dang, V., Croft, W.B.: Diversity by proportionality: an election-based approach to search result diversification. In: SIGIR. pp. 65–74. ACM (2012)
6. Harper, F.M., Konstan, J.A.: The movielens datasets: History and context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5(4), 19 (2016)
7. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: ICDM. pp. 263–272. IEEE (2008)
8. Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: SIGKDD. pp. 426–434. ACM (2008)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS (2013)
10. Radlinski, F., Kleinberg, R., Joachims, T.: Learning diverse rankings with multi-armed bandits. In: ICML. pp. 784–791. ACM (2008)
11. Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: Bpr: Bayesian personalized ranking from implicit feedback. In: UAI. pp. 452–461 (2009)
12. Sakai, T., Song, R.: Evaluating diversified search results using per-intent graded relevance. In: SIGIR. pp. 1043–1052. ACM (2011)
13. Smyth, B., McClave, P.: Similarity vs. diversity. Case-Based Reasoning Research and Development pp. 347–361 (2001)
14. Sun, Y., Yuan, N.J., Wang, Y., Xie, X., McDonald, K., Zhang, R.: Contextual intent tracking for personal assistants. In: SIGKDD. pp. 273–282. ACM (2016)
15. Sun, Y., Yuan, N.J., Xie, X., McDonald, K., Zhang, R.: Collaborative nowcasting for contextual recommendation. In: WWW. pp. 1407–1418 (2016)
16. Sun, Y., Yuan, N.J., Xie, X., McDonald, K., Zhang, R.: Collaborative intent prediction with real-time contextual data. TOIS 35(4), 30 (2017)
17. Tran, T., Phung, D., Venkatesh, S.: Neural choice by elimination via highway networks. In: PAKDD. pp. 15–25. Springer (2016)
18. Vargas, S., Baltrunas, L., Karatzoglou, A., Castells, P.: Coverage, redundancy and size-awareness in genre diversity for recommender systems. In: RecSys (2014)
19. Xia, L., Xu, J., Lan, Y., Guo, J., Cheng, X.: Modeling document novelty with neural tensor network for search result diversification. In: SIGIR (2016)
20. Yuan, M., Pavlidis, Y., Jain, M., Caster, K.: Walmart online grocery personalization: Behavioral insights and basket recommendations. In: International Conference on Conceptual Modeling. pp. 49–64. Springer (2016)
21. Ziegler, C.N., McNee, S.M., Konstan, J.A., Lausen, G.: Improving recommendation lists through topic diversification. In: WWW. pp. 22–32. ACM (2005)