Repeat-bias-aware Optimization of Beyond-accuracy Metrics for Next Basket Recommendation

Yuanna Liu[0000−0002−9868−6578], Ming Li[0000−0001−7430−4961], Mohammad Aliannejadi[0000−0002−9447−4172], and Maarten de Rijke[0000−0002−1086−0202]

University of Amsterdam
{y.liu8,m.li,m.aliannejadi,m.derijke}@uva.nl

Abstract. In next basket recommendation (NBR) a set of items is recommended to users based on their historical basket sequences. In many domains, the recommended baskets consist of both repeat items and explore items. Some state-of-the-art NBR methods are heavily biased to recommend repeat items so as to maximize utility. The evaluation and optimization of beyond-accuracy objectives for NBR, such as item fairness and diversity, has attracted increasing attention. How can such beyond-accuracy objectives be pursued in the presence of heavy repeat bias? We find that only optimizing diversity or item fairness without considering repeat bias may cause NBR algorithms to recommend more repeat items. To solve this problem, we propose a model-agnostic repeat-bias-aware optimization algorithm to post-process the recommended results obtained from NBR methods with the objective of mitigating repeat bias when optimizing diversity or item fairness. We consider multiple variations of our optimization algorithm to cater to multiple NBR methods. Experiments on three real-world grocery shopping datasets show that the proposed algorithms can effectively improve diversity and item fairness, and mitigate repeat bias at acceptable Recall loss.

Keywords: Next basket recommendation · Repeat bias · Beyond-accuracy metrics · Re-ranking

1 Introduction

In next basket recommendation (NBR) a recommender system is meant to recommend a set of items at once [22]. In many e-commerce scenarios in which NBR are deployed users display *repetitive* consumption behavior (e.g., purchasing milk every week, or listening to the same song during workouts) as well as *exploratory* consumption behavior (e.g., buying Christmas gifts) at the same time. Hence, items in a recommended basket can be categorized into repeat items and explore items, depending on whether a user has consumed an item before.

Beyond-accuracy metrics and NBR. Many machine learning techniques have been applied to optimize the accuracy achieved on the NBR task, based on k-nearest neighbors (KNNs) [8, 11, 28], recurrent neural networks (RNNs) [36], or graph neural networks (GNNs) [37]. However, measuring and optimizing beyond-accuracy objectives for NBR remains largely unexplored. Of particular interest are *diversity*, to combat the problem of recommendation homogeneity and generate diversified baskets [15, 33], and

item fairness, to ensure a fair distribution of the exposure assigned to different groups of items [25]. Optimizing for beyond-accuracy metrics such as diversity and item fairness in the context of NBR is made more complex than in other recommendation scenarios due to the presence of repeat bias.

Repeat bias in NBR. Experiments have shown that repeat items contribute most of the accuracy performance [16, 18, 19, 21]. This is because repetition prediction is an easier task, typically with just dozens or hundreds of candidate items and explicit user feedback [20]. However, explore items hold great potential for long-term value in particular and for beyond-accuracy goals in general. E.g., exploring uncertain regions enhances exposure to new and long-tail content, reshaping the overall distribution of the contents, which ultimately improves long-term user experience [32]. Therefore, when curating sets that include both previously seen and new items, it is important to strike the right balance between the repeat and explore categories [2]. Importantly, finding a good balance between repeat and explore items in NBR is made more complex when beyond-accuracy metrics are considered [20]: if most utility is obtained from a relatively small number of items (the repeat items), there is little room to improve beyond-accuracy metrics without sacrificing utility.

Repeat-bias-aware optimization for NBR. For the generic top-k recommendation task, re-ranking is a direct and effective method for improving beyond-accuracy performance, e.g., through greedy algorithms and constrained optimization [39]. The essence of re-ranking is to determine and exploit an effective trade-off between predicted relevance score and beyond-accuracy metrics. However, only optimizing diversity or item fairness may lead to increased repeat bias in some cases according to our experiments, i.e., to an increase in the deviation of the repeat ratio in recommended baskets vs. in ground truth baskets [35]. Even though previous research highlights the problem of repeat bias, no one has tried to optimize to mitigate it. To the best of our knowledge, we are the first to optimize repeat bias jointly with other metrics to seek a balance among (mitigating) repeat bias, accuracy, and beyond-accuracy metrics.

To solve this optimization problem, we propose a model-agnostic repeat-bias-aware diversity optimization (RADiv) algorithm and a repeat-bias-aware item fairness optimization (RAIF) algorithm based on *mixed-integer linear programming* (MILP) for NBR. The repeat ratio (or rather: reducing it) is one of the optimization objectives. The proposed algorithms optimize for predicted relevance, diversity (or item fairness), and repeat ratio simultaneously. We offer several flavors of our MILP-based optimization algorithm, to benefit from the peculiarities of different families of NBR methods. To the best of our knowledge, we are the first to apply a re-ranking algorithm to jointly optimize beyond-accuracy metrics and repeat bias in NBR.

Summarizing, the **main contributions** of the paper are:

- We propose repeat-bias-aware optimization algorithms (RADiv and RAIF) for NBR, which mitigate repeat bias while improving the diversity and item fairness of recommended baskets.
- We extend these algorithms to multiple NBR paradigms, including ones that merge and optimize items from separate repetition and exploration models.

– We conduct experiments on three retail datasets and find that the proposed RADiv and RAIF algorithms can significantly improve diversity, and item fairness and mitigate repeat bias with an acceptable loss in Recall.

2 Related Work

Beyond-accuracy objectives in recommender system. In recommendation, optimizing only for accuracy measures is limiting and misguided [3]. There is a growing interest in beyond-accuracy metrics, which measure other recommendation qualities [30]. Kaminskas and Bridge [13] study the definitions and metrics of diversity, serendipity, novelty, and coverage. They implement re-ranking strategies for these beyond-accuracy metrics to investigate correlations between the two objectives.

Fairness of recommender systems has sparked much research in the community, considering multiple stakeholders of typical e-commerce platforms. From the user side, user fairness requires that a fair system should provide the same recommendation quality for different user groups [23]. From the item side, the goal is to measure the exposure assigned to each item, or each group, and evaluate this distribution to ensure fair principles, such as statistical parity or equal opportunity [29]. Usually, the exposure of an item in a ranked list is computed based on a user browsing model [6, 26].

Beyond-accuracy objectives in NBR. In NBR, most deep-learning-based models predict the top-k relevant items via a user representation to form a basket. This paradigm leads to the problem of over-homogenization of the recommended baskets [33]. To address this problem, Sun et al. [33] apply an autoregressive item-level decoder to generate items one by one to ensure diversified baskets. Leng et al. [15] employ a deconvolutional network to generate diverse NBR results. Regarding item fairness, Liu et al. [25] reproduce a set of item fairness metrics to evaluate representative NBR methods. Li et al. [20] propose a framework to identify short-cuts in achieving better accuracy and beyond-accuracy performance and advocate fine-grained evaluations in NBR.

Apart from the work listed above, optimizing beyond-accuracy objectives in NBR remains unexplored. Regarding the imbalance between repeat and exploration recommendations, Li et al. [19] first define the repeat ratio of the recommended basket and point out the problem of repeat bias in NBR. Furthermore, Tran et al. [35] formulate the repeat bias as the deviation of the repeat ratio of the recommended baskets and ground truth baskets. We contribute by optimizing both diversity and item fairness through re-ranking, while also mitigating repeat bias to improve overall recommendation quality.

Fair/diverse re-ranking. Re-ranking algorithms are usually designed to adjust the ranked results obtained from information access systems considering beyond-accuracy objectives. Commonly used re-ranking algorithms are constrained optimization and maximal marginal relevance (MMR) [5]. For diverse re-ranking, Zhang and Hurley [38] summarize three classic patterns for improving the diversity of recommendation lists as constrained optimization problems: (i) maximize the diversity under the constraint of relevance tolerance; (ii) maximize relevance under the constraint of diversity tolerance; and (iii) maximize the weighted sum of relevance and diversity. In terms of fair re-ranking, Biega et al. [4] propose an online optimization approach that uses integer linear programming (ILP) for re-ranking based on accumulated attention and

relevance scores while constraining according to a bound on the loss of ranking quality. Singh and Joachims [31] optimize a probabilistic ranking to maximize expected utility under three optional fairness constraints. For fair recommendation, Li et al. [23] set the user fairness metric as a constraint to reduce the recommendation quality gap between the advantaged and disadvantaged groups. CPFair [27] simultaneously optimizes consumer and producer fairness in the objective function for multiple rankings. Compared to previous work, which primarily focuses on fairness or diversity in isolation, we are the first to employ a MILP-based re-ranking algorithm to optimize diversity or item fairness, specifically accounting for repeat bias in NBR task.

3 Method

Our notation is summarized in Table 1.

3.1 Problem formulation

In NBR, given a user set U and an item set I, for each user $u \in U$, the purchase history is represented as a sequence of baskets $[B_u^1, B_u^2, \ldots, B_u^t]$, where each basket contains a set of items $B_u^t = \{i_1, i_2, \ldots, i_m | i \in I\}$. In NBR the task is to predict the next basket B_u^{t+1} for each user. Following the common setting, the size of the recommended basket is fixed as K. Generally, NBR methods generate an item ranking $L(u)$ for each user based on the predicted relevance score S_{ui}, and then select the top-K relevant items to form the next basket, i.e., $B_u^{t+1} = L_K(u)$.

3.2 Model overview

In the proposed re-ranking procedure, we select top-N relevant items $L_N(u)$ for each user as candidates. The relevance scores of candidate items of all users can be represented as $S = [S_{ui}]_{|U| \times N}$. A new basket $L'_K(u)$ is selected from the candidate list $L_N(u)$ taking into account the predicted relevance, as well as the diversity, item fairness, and repeat bias of the basket. $L'_K(u) \subset L_N(u), N > K$. The re-ranking procedure can be formalized as a MILP problem, which is a mathematical optimization and can be solved by heuristic algorithms and optimization solvers. In our MILP formulation, the objective function is designed to maximize relevance, diversity (or item fairness) and reduce repeat bias simultaneously. W is a binary decision matrix, $W = [W_{ui}]_{|U| \times N}$. The element $W_{ui} = 1$ indicates recommending item i to user u, and 0 otherwise. The proposed RADiv and RAIF algorithms are specifically adapted for different families of NBR methods: unified methods and combined methods.

3.3 Optimization objectives

In this section, we introduce the definition of three optimization objectives that we aim to achieve.

Diversity. Diverse recommendation aims to recommend items of various and different categories to users. In this work, we use the diversity score (DS) [24] as the diversity

Table 1: Notation used in the paper.

$u \in U$	Users	$i \in I$	Items
$RepRatio_{rec}$	Repeat ratio of recommendation	K	Basket size
$RepRatio_{gt}$	Repeat ratio of ground truth	N	Number of item candidates
W^r	Decision matrix of repeat items	C	Categories of items
W^e	Decision matrix of explore items	I_u^{rep}	Repeat item set of user u
S	Predicted relevance matrix	W	Decision matrix
S^r	Predicted relevance matrix of repeat items	I_1	Popular item group
S^e	Predicted relevance matrix of explore items	I_2	Unpopular item group
$L'_K(u)$	New basket of user u after re-ranking	$L_N(u)$	Item candidate list of user u
$H(\theta)$	Number of repeat items in combined baskets		

objective, computed by dividing the number of recommended categories by the basket size K given the decision matrix W. Diversity objective $\mathrm{DS}(C, W)$ is the sum of DS among all users:

$$\mathrm{DS}(C, W) = \sum_{u \in U} \frac{(\#\text{recommended categories}|W)}{K}. \tag{1}$$

Item fairness. We evaluate the item fairness between popular item group I_1 and unpopular group I_2. The recommended items will receive exposure related to the position. Demographic parity (DP) [31] defines fairness of exposure as equal average exposure between the two groups. The average exposure of a group I_k is computed as:

$$Exposure(I_k \mid W) = \frac{1}{|I_k|} \sum_{i \in I_k} Exposure(i \mid W). \tag{2}$$

Inspired by the construction in [27], the item fairness objective is designed as the difference of average exposure between group I_1 and I_2. The closer the IF value is to zero, the fairer a ranking is.

$$\mathrm{IF}(I_1, I_2, W) = Exposure(I_1 \mid W) - Exposure(I_2 \mid W). \tag{3}$$

Repeat bias. Some NBR methods are either biased to recommending too many repeat items or explore items compared with the ground truth baskets. We aim to mitigate the repeat bias of recommended baskets. Since the repeat ratio of ground truth baskets $RepRatio_{gt}$ is unknown, we directly choose the repeat ratio of recommended baskets $RepRatio_{rec}$ as the optimization objective. Each user has a repeat item set I_u^{rep}, which contains all the items the user has bought before. Given the decision matrix W, the repeat ratio objective can be expressed as:

$$RepRatio_{rec}(I_u^{rep}, W) = \sum_{u \in U} \frac{(|L'_K(u) \cap I_u^{rep}| \,|W)}{K}. \tag{4}$$

3.4 Repeat-bias-aware optimization for unified NBR methods

In this section, we post-process the recommendation lists of users generated by unified NBR methods, where a unified model generates the relevance scores of all items. We select top-N relevant items of each user as candidates. The relevance scores of these

item candidates can be represented as $S = [S_{ui}]_{|U| \times N}$. We apply a MILP model to re-rank the top N candidates for each user. RADiv algorithm simultaneously optimizes diversity and repeat ratio as shown in Eq. 5:

$$\begin{aligned} &\max \frac{1}{K} \sum_{u \in U} \sum_{i=1}^{N} S_{ui} W_{ui} + \epsilon_1 \, \mathrm{DS}(C, W) - \lambda RepRatio_{rec}(I_u^{rep}, W) \\ &\text{such that } \sum_{i=1}^{N} W_{ui} = K, W_{ui} \in \{0, 1\}. \end{aligned} \tag{5}$$

We design the objective function to maximize the sum of relevance scores, the diversity score $\mathrm{DS}(C, W)$, while minimizing the $RepRatio_{rec}(I_u^{rep}, W)$. Here, ϵ_1 and λ are the weighting parameters of the diversity term and repeat ratio term, respectively. W is a binary decision matrix determining whether to recommend item i to user u. The constraint indicates that the algorithm ultimately recommends K items to each user, which is the basket size. It is worth noting that the minus sign in front of $RepRatio_{rec}$ is designed for repeat-biased NBR methods with quite high $RepRatio_{rec}$. The minus implies that the algorithm aims to reduce the $RepRatio_{rec}$ to be closer to $RepRatio_{gt}$. In contrast, for explore-biased methods with quite low $RepRatio_{rec}$ values, this term should be set to a positive sign: $+\lambda RepRatio_{rec}(I_u^{rep}, W)$. The algorithm will increase the $RepRatio_{rec}$ so as to approach $RepRatio_{gt}$.

Similarly, RAIF simultaneously optimizes item fairness and repeat ratio by Eq. 6.

$$\begin{aligned} &\max \sum_{u \in U} \sum_{i=1}^{N} S_{ui} W_{ui} - \alpha_1 \, \mathrm{IF}(I_1, I_2, W) - \lambda RepRatio_{rec}(I_u^{rep}, W) \\ &\text{such that } \sum_{i=1}^{N} W_{ui} = K, W_{ui} \in \{0, 1\}. \end{aligned} \tag{6}$$

The objective function is designed to maximize the combination of relevance scores, item fairness $\mathrm{IF}(I_1, I_2, W)$, and minimize $RepRatio_{rec}(I_u^{rep}, W)$. Here, α_1 and λ are the weighting parameters of item fairness term and repeat ratio term, respectively.

3.5 Repeat-bias-aware optimization for combined NBR methods

In the previous setting, the optimization algorithm is designed for unified NBR methods, where the relevance scores of repeat items and explore items are generated by the same model and are comparable. However, there is another NBR paradigm, such as two-step repetition-exploration (TREx) [20], where the repeat item list and explore item list are obtained from different models. This paradigm allows one to combine the strongest repetition model and the strongest exploration model flexibly, and adjust the proportion of these two parts. When we optimize diversity (or item fairness) and repeat bias of combined NBR methods, the challenge lies in the fact that the predicted relevance scores are not comparable between repeat items and explore items.

Inspired by the TREx framework [20], we use a threshold θ to filter the repeat item candidates and count the number of repeat items with a relevance score larger than θ.

Algorithm 1: RADiv and RAIF for Combined NBR Methods

Input: Basket size K, Categories C, Number of item candidates N, Item group I_1, I_2, Parameters $\epsilon_2, \alpha_2, \theta$

Output: Recommendation matrix W^r, W^e

$S^r, S^e \leftarrow$ The top-N repeat and explore relevance scores of users.

Compare θ with S^r_{ui}, and obtain $H(\theta) = \min(H(\theta)_{\text{aux}}, K)$.

Solve the optimization problem following Eq. 7 (or 8).

Return W^r, W^e

The numbers of repeat items for each user are saved as vector $H(\theta)_{\text{aux}}$. The final repeat position number for each user is $H(\theta) = \min(H(\theta)_{\text{aux}}, K)$. Here, θ indirectly controls $RepRatio_{rec}(\theta)$. The higher the θ becomes, the lower the number of repeat slots $H(\theta)$ and $RepRatio_{rec}(\theta)$ become. $H(\theta)$ is an important variable to avoid the comparison between repeat and explore relevance scores. In this way, repeat items are compared internally and $H(\theta)$ of them are finally selected. Explore items compete internally and $K - H(\theta)$ of them are chosen.

The RADiv algorithm is adapted for combined NBR methods as shown in Eq. 7. $S^r = [S^r_{ui}]_{|U|\times N}$ and $S^e = [S^e_{ui}]_{|U|\times N}$ are predicted relevance score matrices of repeat items and explore items from different recommender systems, respectively. $W^r = [W^r_{ui}]_{|U|\times N}$, $W^e = [W^e_{ui}]_{|U|\times N}$ are corresponding binary decision matrices to determine the selection of repeat items and explore items. ϵ_2 is the weighting parameter of the diversity term. The constraints indicate the repeat slots for each user are $H(\theta)$, while the explore slots for each user are $K - H(\theta)$. Similarly, RAIF is adjusted to optimize item fairness $\text{IF}(I_1, I_2, W^r, W^e)$ and $RepRatio_{rec}(\theta)$ in Eq. 8. α_2 is the weighting parameter of item fairness term. Algorithm 1 summarizes our algorithms for combined NBR methods.

$$\begin{aligned} &\max \frac{1}{K}\sum_{u\in U}\sum_{i=1}^{N}(S^r_{ui}W^r_{ui} + S^e_{ui}W^e_{ui}) + \epsilon_2\,\text{DS}(C, W^r, W^e) \\ &\quad \text{such that } \sum_{i=1}^{N} W^r_{ui} = H(\theta), \sum_{i=1}^{N} W^e_{ui} = K - H(\theta), W^r_{ui}, W^e_{ui} \in \{0,1\}; \end{aligned} \tag{7}$$

$$\begin{aligned} &\max \sum_{u\in U}\sum_{i=1}^{N}(S^r_{ui}W^r_{ui} + S^e_{ui}W^e_{ui}) - \alpha_2\,\text{IF}(I_1, I_2, W^r, W^e) \\ &\quad \text{such that } \sum_{i=1}^{N} W^r_{ui} = H(\theta), \sum_{i=1}^{N} W^e_{ui} = K - H(\theta), W^r_{ui}, W^e_{ui} \in \{0,1\}. \end{aligned} \tag{8}$$

4 Experimental Setup

NBR methods. We select and investigate the following 5 representative NBR methods:

– **UP-CF@r**, which combines recency-aware user-wise popularity and collaborative filtering while considering the recent shopping behavior [8].

Table 2: Statistics of the datasets after preprocessing.

Dataset	#Users	#Items	#Baskets	Avg. #baskets/user	Avg. #items/basket	Avg. $RepRatio_{gt}$
Instacart	19,210	29,399	305,582	15.91	10.06	0.60
Dunnhumby	2,482	37,162	107,152	43.17	10.07	0.43
TaFeng	10,182	15,024	82,387	8.09	6.14	0.21

– **TIFUKNN**, which models the temporal dynamics of users' past baskets by using a KNN-based approach based on the personalized item frequency (PIF) [11].
– **Dream**, which forms basket representations using a pooling strategy and models sequential user behavior through an RNN [36].
– **DNNTSP** uses GNN and self-attention mechanisms to encode item-to-item relations across baskets and capture temporal dependencies [37].
– **TREx**, which has a repetition module considering item repurchase features and users' interests, and an exploration module targeted for beyond-accuracy metrics. Then, repeat items and explore items are combined to form the final basket [20].

We exclude NBR methods that only focus on the repetition recommendation (i.e., P-TopFreq, ReCANet [1], and NBRR [14]) or exploration recommendation (i.e., NNBR [17]), as we aim to investigate and balance repetition *and* exploration in NBR.

Datasets. Following previous NBR studies [17, 19, 20, 25], we select three publicly available grocery shopping datasets: TaFeng [34], Dunnhumby [7], and Instacart [12]. They exhibit different characteristics in terms of repetition and exploration, which are critical for verifying the effectiveness of our proposed repeat-bias-aware optimization algorithms.

For each dataset, we remove users with fewer than three baskets and items purchased fewer than five times, as done in [1]. Due to the large size of the Instacart dataset, memory limitations occurred during the calculation of some methods. Following [28], we also randomly sampled 20,000 users from Instacart before applying any filtering. Table 2 provides the statistics of the three datasets after preprocessing. The average $RepRatio_{gt}$ refers to the average proportion of repeat items in the ground truth baskets, as defined in [19].

We split each dataset following the approach in [1, 8, 28]. The training set includes all baskets from each user except the last basket. For users with more than 50 baskets in the training data, we limit the training set to their most recent 50 baskets [19]. The last baskets of all users are then randomly split equally into a validation set (50%) and a test set (50%).

Evaluation metrics. We use the following widely used metrics in our experiments to measure how the proposed algorithms balance multiple objectives. In terms of accuracy, Recall measures the system's ability to retrieve items that users will purchase in their next baskets. For item fairness, demographic parity (DP) measures the ratio of the average exposure of the popular group to the average exposure of the unpopular group. Following [29], we use logDP to deal with the empty-group case. The closer the logDP value is to zero, the fairer the recommendation is. The diversity score (DS) measures the number of categories within the recommended basket divided by basket size. Re-

peat bias is computed as: RepBias = $RepRatio_{rec} - RepRatio_{gt}$ [35]. Following [27], we use a comprehensive metric to evaluate the overall performance of item fairness and RepBias: mFR = $\omega|\text{logDP}| + (1-\omega)|\text{RepBias}|$. Similarly, the overall performance of diversity and RepBias is evaluated as mDR = $\omega\text{DS} - (1-\omega)|\text{RepBias}|$.

Implementation details. In our experiments, we follow [11, 20] and set basket size $K = 20$. The number of item candidates $N = 100$. We compute the popularity of all items and select the top 20% as popular I_1 and the rest belong to the unpopular group I_2. We apply grid search to select hyperparameters on the validation set. We select α_1 and α_2 in [0, 0.001, 0.01, 0.1, 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200]. ϵ_1 and select ϵ_2 from [0, 0.001, 0.01, 0.02, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18, 0.2]. We choose λ in [0, 0.001, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]. θ candidates are selected based on S^r.

We select the optimal hyperparameter combination according to these two conditions: (i) satisfy a 10% Recall drop tolerance; and (ii) maximize mDR or minimize mFR, where $\omega = 0.5$ following [27]. We use Gurobi,[1] which is an industrial optimization solver capable of delivering practical and effective feasible solutions. We release our code and hyperparameters at https://github.com/lynEcho/Repbias_NBR.

5 Experimental Results

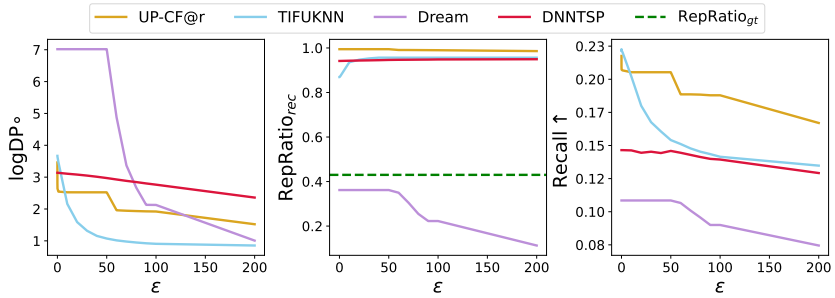
We design experiments to answer three research questions: **RQ1** How does optimizing only for diversity or item fairness affect the repeat bias? **RQ2** What is the performance of proposed RADiv and RAIF algorithms on NBR methods? **RQ3** How does RADiv and RAIF algorithms strike a balance between utility and beyond-accuracy metrics?

We start by studying how optimizing for beyond-accuracy metrics, without taking the repeat bias, would affect the baskets in terms of repeat-explore items, answering **RQ1**. We apply a re-ranking algorithm without considering $RepRatio_{rec}$ to optimize the recommended baskets obtained from different NBR methods on Instacart, Dunnhumby, and TaFeng. For diverse re-ranking, the optimization objective is formulated as $\max \frac{1}{K} \sum_{u \in U} \sum_{i=1}^{N} S_{ui} W_{ui} + \epsilon\, \text{DS}(C, W)$. In terms of fair re-ranking, the objective function is $\max \sum_{u \in U} \sum_{i=1}^{N} S_{ui} W_{ui} - \alpha\, \text{IF}(I_1, I_2, W)$. The parameters ϵ and α are used to adjust the weight of the diversity and item fairness terms, respectively. Fig. 1 shows the performance of diversity and item fairness optimization on Dunnhumby.[2]

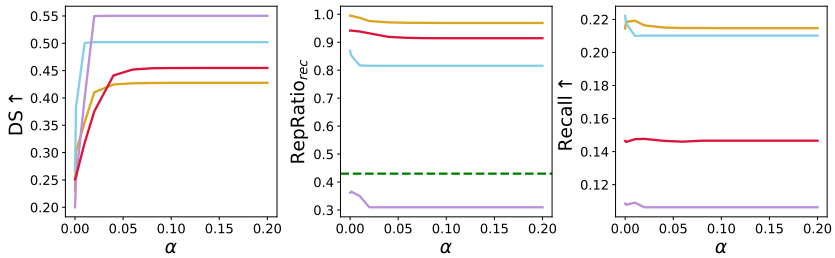
We make the following observation: (i) Diversity and item fairness optimizations significantly improve the NBR methods in terms of logDP and DS. For item fairness optimization, as the parameter ϵ increases ($0 \rightarrow 200$), all NBR methods become fairer. Also, the DS of all NBR methods increase as α increases ($0 \rightarrow 0.2$). (ii) In most cases, Recall exhibits a declining trend as DS and logDP become better. In diversity optimization, the Recall of UP-CF@r, Dream, and DNNTSP improve slightly when $\alpha = 0.01$. (iii) In terms of repeat bias, which is measured as the gap between $RepRatio_{rec}$ and $RepRatio_{gt}$ (green dashed line in Fig. 1). We see that the repeat bias is likely to either intensify or diminish, showing no clear trend. In item fairness optimization, the

[1] https://www.gurobi.com

[2] We observe a similar trend on the Instacart and TaFeng datasets. Because of space limitations, we report the results on Instacart and TaFeng in the anonymous repository.



(a) Item fairness optimization



(b) Diversity optimization

Fig. 1: Performance of item fairness and diversity optimization without considering $RepRatio_{rec}$ on different NBR methods. logDP∘ means the closer logDP value is to zero, the fairer it is.

repeat bias of TIFUKNN and Dream grows as ϵ increases, showing that it is important to consider repeat bias while optimizing beyond-accuracy metrics.

To answer **RQ2**, we apply RADiv and RAIF algorithms on five NBR methods. Of these, TIFUKNN, UP-CF@r, DNNTSP, and Dream are unified repeat-explore recommendation methods while TREx is a combined method. Since TREx does not provide predicted relevance scores for explore items, here we use the explore item list obtained from UP-CF@r to combine with repeat items of TREx. Table 3 reports the results on Instacart.[3] We also conduct an ablation study, i.e., optimizing only for diversity, item fairness, and repeat bias, respectively.

We arrive at the following conclusions: (i) The proposed RADiv algorithm (RD) achieves a significant improvement in DS, RepBias, and comprehensive metric mDR in comparison to the original (Ori.) results on all NBR methods. Surprisingly, the Recall of Dream increases after RADiv optimization ($0.0977 \rightarrow 0.1468$). (ii) The RAIF algorithm (RF) achieves better logDP, less RepBias, and better overall metric mFR than the original (Ori.) results, proving the effectiveness of RAIF algorithm. (iii) RADiv and RAIF algorithms can effectively adjust the $RepRatio_{rec}$ of methods to approach $RepRatio_{gt}$. Take Instacart as an example, RADiv and RAIF algorithms decrease the $RepRatio_{rec}$ of TIFUKNN, UP-CF@r, DNNTSP, and increase the $RepRatio_{rec}$ of Dream to make them closer to ground truth 0.6. (iv) In our ablation study, we find that

[3] Similar patterns are observed on the Dunnhumby and TaFeng datasets. Due to space limitations, we report the results on Dunnhumby and TaFeng in the anonymous repository.

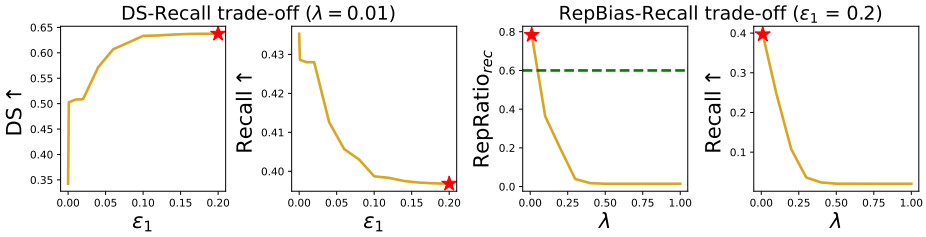
Table 3: The performance of RADiv and RAIF algorithms on Instacart ($RepRatio_{gt} = 0.6$). *Ori.* indicates the original baskets obtained from each method. *RD* indicates RADiv. *RF* refers to RAIF. In ablation study, *D*, *F*, *R* indicate optimizing only for diversity, item fairness, and repeat bias, respectively. RepR indicates $RepRatio_{rec}$ (best close to $RepRatio_{gt}$). RepBias and logDP (best close to zero).

Meth.	Diversity optimization						Item fairness optimization					
	Type	Recall↑	DS↑	RepR	RepBias	mDR↑	Type	Recall↑	logDP	RepR	RepBias	mFR↓
TIFUKNN	Ori.	**0.4559**	0.3615	0.9248	0.3248	0.0184	Ori.	**0.4559**	3.1252	0.9248	0.3248	1.7250
	D	0.4259	0.5897	0.8984	0.2984	0.1457	F	0.4269	2.3510	0.9252	0.3252	1.3381
	R	0.4537	0.3587	0.9100	0.3100	0.0244	R	0.4232	3.2760	**0.7939**	**0.1939**	1.7350
	RD	0.4245	**0.5898**	**0.8874**	**0.2874**	**0.1512**	RF	0.4098	**2.3271**	0.8718	0.2718	**1.2995**
UP-CF@r	Ori.	**0.4405**	0.3489	0.8905	0.2905	0.0292	Ori.	**0.4405**	3.3966	0.8905	0.2905	1.8436
	D	0.3983	**0.6375**	0.7896	0.1896	0.2239	F	0.4282	2.4860	0.8858	0.2858	1.3859
	R	0.4353	0.3424	0.8812	0.2812	0.0306	R	0.4373	3.3997	0.8791	0.2791	1.8394
	RD	0.3968	**0.6375**	**0.7837**	**0.1837**	**0.2269**	RF	0.4052	**2.1810**	**0.8019**	**0.2019**	**1.1915**
DNNTSP	Ori.	**0.4347**	0.3402	0.9133	0.3133	0.0135	Ori.	**0.4347**	3.2573	0.9133	0.3133	1.7853
	D	0.4046	0.6009	0.8660	0.2660	0.1675	F	0.4332	**2.9780**	0.9152	0.3152	1.6466
	R	0.4337	0.3388	0.9080	0.3080	0.0154	R	0.4256	3.3204	**0.8640**	**0.2640**	1.7922
	RD	0.4059	**0.6013**	**0.8623**	**0.2623**	**0.1695**	RF	0.4277	2.9901	0.8837	0.2837	**1.6369**
Dream	Ori.	0.0977	0.1000	0.1923	-0.4077	-0.1539	Ori.	0.0977	7.3111	0.1923	-0.4077	3.8594
	D	0.0723	**0.7000**	0.1267	-0.4733	0.1133	F	0.0709	**2.1018**	0.1226	-0.4774	**1.2896**
	R	**0.1540**	0.1243	**0.4202**	**-0.1798**	-0.0277	R	**0.1288**	7.3111	**0.3108**	**-0.2892**	3.8002
	RD	0.1468	0.5654	**0.4202**	**-0.1798**	**0.1928**	RF	0.0939	2.3893	0.2094	-0.3906	1.3900
TREx	Ori.	**0.4595**	0.3533	0.9265	0.3265	0.0134	Ori.	**0.4595**	3.2182	0.9265	0.3265	1.7724
	D	0.4455	0.5296	0.9265	0.3265	0.1016	F	0.4569	2.9052	0.9265	0.3265	1.6159
	R	0.4368	0.3280	**0.7973**	**0.1973**	0.0654	R	0.4277	3.4281	**0.7487**	**0.1487**	1.7884
	RD	0.4195	**0.5874**	**0.7973**	**0.1973**	**0.1951**	RF	0.4226	**2.6528**	**0.7487**	**0.1487**	**1.4008**

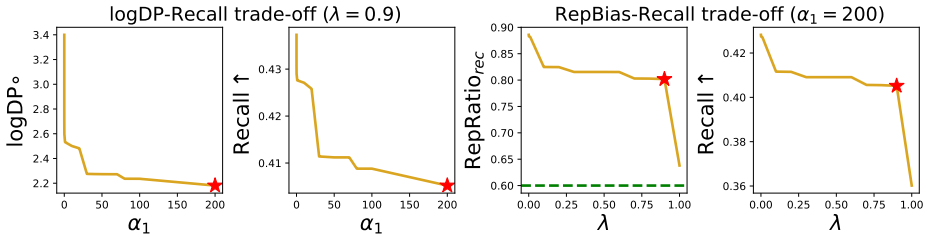
only optimizing $RepRatio_{rec}$ (R) will decrease the diversity and item fairness of TIFUKNN, UC-CF@r, DNNTSP and TREx. This means that it is necessary to optimize both $RepRatio_{rec}$ and diversity, both $RepRatio_{rec}$ and item fairness, as implemented by the proposed RADiv and RAIF algorithms.

To answer **RQ3**, we take UP-CF@r as an example to illustrate the trade-off between Recall and other metrics on Instacart. For the two parameters ϵ_1 and λ in the RADiv optimization, we change one parameter each time with the other one fixed. Similarly, we perform the same operation for α_1 and λ in RAIF optimization. The analysis is shown in Fig. 2.

We make the following observations: (i) In Fig. 2a, we first fix λ and change ϵ_1 from 0 to 0.2. We see that DS continuously increases at the cost of Recall. The red star indicates that the algorithm chooses $\epsilon_1 = 0.2$ as the optimal solution to achieve a balance between diversity and Recall. When we fix ϵ_1 and change λ from 0 to 1, $RepRatio_{rec}$ and Recall decline consistently. Since our algorithm optimizes $RepRatio_{rec}$ to be approaching $RepRatio_{gt}$, it selects $\lambda = 0.01$ as the optimal trade-off between Recall and repeat bias. This process reflects the balanced strategy of RADiv to maximize diversity



(a) RADiv optimization



(b) RAIF optimization

Fig. 2: Trade-off between Recall and diversity, item fairness, repeat bias. Take UP-CF@r as an example on the Instacart dataset. The red star indicates the optimal solution chosen by the proposed algorithm. The green dashed line is value of $RepRatio_{gt}$.

and mitigate repeat bias within the tolerance of utility loss. (ii) In Fig. 2b, we fix λ and adjust α_1 from 0 to 200. logDP decreases indicating unpopular items are assigned with more exposure. The recommendation becomes fairer at the cost of a Recall drop. The algorithm chooses $\alpha_1 = 200$ to achieve a balance between logDP and Recall. When we change λ from 0 to 1 with a fixed α_1, $RepRatio_{rec}$ and Recall show a consistent downward trend. The algorithm selects $\lambda = 0.9$ to balance repeat bias and Recall. Considering the inverse relationship between Recall and logDP, repeat bias, the RAIF algorithm decides to sacrifice a little utility in exchange for a greater return on item fairness and repeat bias.

6 Conclusion

We have proposed repeat-bias-aware optimization algorithms for improving diversity and item fairness while mitigating repeat bias in NBR. Our RADiv and RAIF algorithms re-rank the preliminary recommended results obtained from various NBR baselines according to different beyond-accuracy objectives and seek a balance between repeat items and explore items simultaneously while optimizing for beyond-accuracy metrics. We have extended our approach to multiple NBR paradigms [20], in particular, one that fuses repeat item lists and explores item lists from separate models, allowing us to select tailored models for each list. We have conducted experiments on three real-world retail datasets. We find that only optimizing for diversity or item fairness will increase repeat bias in many cases, which can reduce user satisfaction. The proposed RADiv and RAIF algorithms can effectively optimize diversity and fairness while mitigating the repeat bias issue under an acceptable utility loss. Finally, we have investigated

trade-offs between Recall and other beyond-accuracy metrics, including diversity, item fairness, and repeat bias.

Based on our experiments, we find that it is critical to evaluate the utility of NBR methods and measure their repeat bias at the same time. In many cases, a recommendation model can achieve the best performance in terms of Recall by recommending only repeat items (i.e., highest repeat bias), but this reduces the likelihood of novelty and serendipity in recommendations, leading to filter bubble and echo chamber issues [9, 10]. Re-ranking may be an effective solution to this problem, as we have shown that it can strike a balance between accuracy and beyond-accuracy metrics, while taking into account the repeat bias in the re-ranking process. However, in extreme cases where the original ranking does not include many explore items, re-ranking cannot really be effective. Therefore, we suggest to consider accuracy and beyond-accuracy objectives in the original ranking. In addition, the definition and quantification of repeat bias is also a direction worth exploring in the future.

Acknowledgments

This research was (partially) supported by the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union's Horizon Europe program under grant agreement No 101070212.

All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Bibliography

[1] Ariannezhad, M., Jullien, S., Li, M., Fang, M., Schelter, S., de Rijke, M.: RecANet: A repeat consumption-aware neural network for next basket recommendation in grocery shopping. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1240–1250 (2022)

[2] Ariannezhad, M., Li, M., Jullien, S., de Rijke, M.: Complex item set recommendation. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3444–3447 (2023)

[3] Bauer, C., Bagchi, C., Hundogan, O.A., van Es, K.: Where are the values? A systematic literature review on news recommender systems. ACM Transactions on Recommender Systems (2024)

[4] Biega, A.J., Gummadi, K.P., Weikum, G.: Equity of attention: Amortizing individual fairness in rankings. In: The 41st international acm sigir conference on research & development in information retrieval, pp. 405–414 (2018)

[5] Carbonell, J., Goldstein, J.: The use of mmr, diversity-based reranking for reordering documents and producing summaries. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 335–336 (1998)

[6] Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P.: Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM conference on Information and knowledge management, pp. 621–630 (2009)

[7] Dunnhumby: Source files (2024), https://www.dunnhumby.com/source-files/

[8] Faggioli, G., Polato, M., Aiolli, F.: Recency aware collaborative filtering for next basket recommendation. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization, pp. 80–87 (2020)

[9] Flaxman, S., Goel, S., Rao, J.M.: Filter bubbles, echo chambers, and online news consumption. Public opinion quarterly **80**, 298–320 (2016)

[10] Ge, Y., Zhao, S., Zhou, H., Pei, C., Sun, F., Ou, W., Zhang, Y.: Understanding echo chambers in e-commerce recommender systems. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2261–2270 (2020)

[11] Hu, H., He, X., Gao, J., Zhang, Z.L.: Modeling personalized item frequency information for next-basket recommendation. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1071–1080 (2020)

[12] Instacart: Market basket analysis (2017), https://www.kaggle.com/c/instacart-market-basket-analysis/data

[13] Kaminskas, M., Bridge, D.: Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Transactions on Interactive Intelligent Systems (TiiS) **7**(1), 1–42 (2016)

[14] Katz, O., Barkan, O., Koenigstein, N., Zabari, N.: Learning to ride a buy-cycle: A hyper-convolutional model for next basket repurchase recommendation. In: Pro-

ceedings of the 16th ACM Conference on Recommender Systems, pp. 316–326 (2022)
[15] Leng, Y., Yu, L., Xiong, J., Xu, G.: Recurrent convolution basket map for diversity next-basket recommendation. In: International conference on database systems for advanced applications, pp. 638–653, Springer (2020)
[16] Li, M., Ariannezhad, M., Yates, A., De Rijke, M.: Who will purchase this item next? Reverse next period recommendation in grocery shopping. ACM Transactions on Recommender Systems **1**(2), 1–32 (2023)
[17] Li, M., Ariannezhad, M., Yates, A., de Rijke, M.: Masked and swapped sequence modeling for next novel basket recommendation in grocery shopping. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 35–46 (2023)
[18] Li, M., Huang, J., de Rijke, M.: Repetition and exploration in offline reinforcement learning-based recommendations. In: Proceedings of the 4th Workshop on Deep Reinforcement Learning for Information Retrieval at CIKM, ACM (October 2023)
[19] Li, M., Jullien, S., Ariannezhad, M., de Rijke, M.: A next basket recommendation reality check. ACM Transactions on Information Systems **41**(4), 1–29 (2023)
[20] Li, M., Liu, Y., Jullien, S., Ariannezhad, M., Yates, A., Aliannejadi, M., de Rijke, M.: Are we really achieving better beyond-accuracy performance in next basket recommendation? In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 924–934 (2024)
[21] Li, M., Vardasbi, A., Yates, A., de Rijke, M.: Repetition and exploration in sequential recommendation. In: Proceedings of the 46th international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2532–2541, ACM (July 2023)
[22] Li, R., Zhang, L., Liu, G., Wu, J.: Next basket recommendation with intent-aware hypergraph adversarial network. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1303–1312 (2023)
[23] Li, Y., Chen, H., Fu, Z., Ge, Y., Zhang, Y.: User-oriented fairness in recommendation. In: Proceedings of the web conference 2021, pp. 624–632 (2021)
[24] Liang, Y., Qian, T., Li, Q., Yin, H.: Enhancing domain-level and user-level adaptivity in diversified recommendation. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 747–756 (2021)
[25] Liu, Y., Li, M., Ariannezhad, M., Mansoury, M., Aliannejadi, M., de Rijke, M.: Measuring item fairness in next basket recommendation: A reproducibility study. In: European Conference on Information Retrieval, pp. 210–225, Springer (2024)
[26] Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Transactions on Information Systems (TOIS) **27**(1), 1–27 (2008)
[27] Naghiaei, M., Rahmani, H.A., Deldjoo, Y.: CPFair: Personalized consumer and producer fairness re-ranking for recommender systems. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 770–779 (2022)
[28] Naumov, S., Ananyeva, M., Lashinin, O., Kolesnikov, S., Ignatov, D.I.: Time-dependent next-basket recommendations. In: European Conference on Information Retrieval, pp. 502–511, Springer (2023)

[29] Raj, A., Ekstrand, M.D.: Measuring fairness in ranked results: An analytical and empirical comparison. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 726–736 (2022)
[30] de Rijke, M.: Beyond accuracy goals, again. In: WSDM 2023: The Sixteenth International Conference on Web Search and Data Mining, pp. 2–3, ACM (2023)
[31] Singh, A., Joachims, T.: Fairness of exposure in rankings. In: Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pp. 2219–2228 (2018)
[32] Su, Y., Wang, X., Le, E.Y., Liu, L., Li, Y., Lu, H., Lipshitz, B., Badam, S., Heldt, L., Bi, S., et al.: Long-term value of exploration: Measurements, findings and algorithms. In: Proceedings of the 17th ACM International Conference on Web Search and Data Mining, pp. 636–644 (2024)
[33] Sun, W., Xie, R., Zhang, J., Zhao, W.X., Lin, L., Wen, J.R.: Generative next-basket recommendation. In: Proceedings of the 17th ACM Conference on Recommender Systems, pp. 737–743 (2023)
[34] TaFeng: Grocery dataset (2001), https://www.kaggle.com/datasets/chiranjivdas09/ta-feng-grocery-dataset
[35] Tran, V.A., Salha-Galvan, G., Sguerra, B., Hennequin, R.: Transformers Meet ACT-R: Repeat-Aware and Sequential Listening Session Recommendation. RecSys2024 (2024)
[36] Yu, F., Liu, Q., Wu, S., Wang, L., Tan, T.: A dynamic recurrent model for next basket recommendation. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, pp. 729–732 (2016)
[37] Yu, L., Sun, L., Du, B., Liu, C., Xiong, H., Lv, W.: Predicting temporal sets with deep neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1083–1091 (2020)
[38] Zhang, M., Hurley, N.: Avoiding monotony: improving the diversity of recommendation lists. In: Proceedings of the 2008 ACM conference on Recommender systems, pp. 123–130 (2008)
[39] Zhao, Y., Wang, Y., Liu, Y., Cheng, X., Aggarwal, C.C., Derr, T.: Fairness and diversity in recommender systems: a survey. ACM Transactions on Intelligent Systems and Technology (2024)