Repetition and Exploration in Sequential Recommendation

Ming Li
University of Amsterdam
Amsterdam, The Netherlands
m.li@uva.nl

Ali Vardasbi
University of Amsterdam
Amsterdam, The Netherlands
a.vardasbi@uva.nl

Andrew Yates
University of Amsterdam
Amsterdam, The Netherlands
a.c.yates@uva.nl

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

In several recommendation scenarios, including next basket recommendation, the importance of repetition and exploration has been discovered and studied. Sequential recommenders (SR) aim to infer a user's preferences and suggest the next item for them to interact with based on their historical interaction sequences. There has not been a systematic analysis of sequential recommenders from the perspective of repetition and exploration. As a result, it is unclear how these models, that are typically optimized for accuracy, perform in terms of repetition and exploration, as well as the potential drawbacks of deploying them in real applications.

In this paper, we examine whether repetition and exploration are important dimensions in the sequential recommendation scenario. We consider this generalizability question both from a user-centered and an item-centered perspective. Towards the latter, we define *item repeat exposure* and *item explore exposure* and examine the recommendation performance of sequential recommendation models in terms of both accuracy and exposure from the perspective of repetition and exploration. We find that (i) there is an imbalance in accuracy and difficulty w.r.t. repetition and exploration in SR scenarios, (ii) using the conventional average overall accuracy with a significance test does not fully represent a model's recommendation accuracy, and (iii) accuracy-oriented sequential recommendation models may suffer from less/zero item explore exposure issue, where items are mostly (or even only) recommended to their repeat users and fail to reach their potential new users.

To analyze our findings, we remove repeat samples from the dataset, which often act as easy shortcuts, and focus on a pure exploration SR scenario. We find that (i) removing the repetition shortcut increases recommendation novelty and helps users who prefer to consume novel items next, (ii) neural-based models fail to learn the basic characteristics of this pure exploration scenario and suffer from an inherent repetitive bias issue, (iii) using shared item embeddings in the prediction layer may skew recommendations to repeat items, and (iv) removing all repeat items by post-processing recommendation results leads to a substantial improvement on top of several SR methods.

CCS CONCEPTS

• **Information systems → Recommender systems**; *Retrieval models and ranking*.

KEYWORDS

Sequential recommendation; Repetition and exploration; Explore exposure

ACM Reference Format:
Ming Li, Ali Vardasbi, Andrew Yates, and Maarten de Rijke . 2023. Repetition and Exploration in Sequential Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3539618.3591914

1 INTRODUCTION

Recommender systems have become an essential instrument for connecting users and items on many online platforms [44]. Users may have dynamic interests over time, and sequential recommender systems aim to learn from users' historical interaction sequences to infer their preferences and suggest an appropriate next item for them to interact with [9, 27, 34]. Advances in deep learning have led to the development of numerous sequential recommendation models that employ deep learning techniques such as RNNs [12, 13], CNNs [31], GNNs [29, 37], contrastive learning [39], attention mechanisms [19, 21], and self-attention [16, 30, 33].

Repetition and exploration. The default focus of sequential recommendation is on increasing accuracy, that is, to find relevant or correct next items that meet the preferences of users. In the next basket recommendation (NBR) scenario, Li et al. [20] distinguish between *repetition*, i.e., when the next item the user interacts with is present in the user's historical interaction sequence, and *exploration*, i.e., when the user first interacts with an item they have not previously interacted with. The authors find very large differences in performance when recommending repeat items vs. explore items, with the task of recommending repeat items being far easier and achieving far higher accuracy scores. As repetition and exploration behavior coexist in many sequential recommendation scenarios, such as item repurchase [3, 5], song relistening [1], and POI revisits [7], a natural question to ask is:

SIGIR '23, July 23–27, 2023, Taipei, Taiwan

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591914

> *How do sequential recommendation models perform from the repetition and exploration perspective?*

Sequential recommendation: a user-centered perspective. To address the question highlighted above, we first adopt a user-centered perspective. We select a diverse set of highly-cited sequential recommendation models [13, 16, 28, 30, 31, 37] to examine if a similar imbalance between repetition performance and exploration performance as was found for NBR is also observed from the point of the users who are being served sequential recommendations. We consider the case where each user purchases one item and that item can either be a *repeat item* (an item the user has bought before) or an *explore item* (an item the user has not purchased before).

We find that users who prefer repetition over exploration get noticeably higher recommendation accuracy than users who prefer to explore. We also find that a higher overall accuracy (aggregated over all users) can be achieved by sacrificing the performance for users who prefer to explore.

These findings matter because the average overall accuracy can be achieved by sacrificing the quality of recommendations for a large proportion of users, which challenges the widely adopted usage of average accuracy with significance test in SR research for evaluation.

Sequential recommendation: an item-centered perspective. The user-centered evaluation summarized above only provides a partial perspective on the capabilities of a recommendation algorithm. There is at least one more side to the (sequential) recommendation task: items. *Item exposure* refers to how often an item is recommended by a recommendation algorithm. Item exposure can have a significant impact on the user experience and the overall effectiveness of the system [8, 10].

Previous studies regarding item exposure often focus on fairness [36, 38]. In this paper, we generalize the highlighted question in two ways: from the next basket recommendation scenario to the sequential recommendation scenario and also from a user-centered perspective to an item-centered perspective. Specifically, we distinguish between *item repeat exposure* and *item explore exposure*: The former refers to the number of times the item is exposed to repeat users, i.e., users who have purchased it before, whereas the latter refers to the number of times an item gets exposed to new users, i.e., users who have not purchased it before. The motivation for this perspective is that if an item has been purchased by a large proportion of new users in the future, then this item should probably be recommended to many new users.

We first analyze the distribution of items' next target users (in historical interaction logs) and observe that for most items, there exists a large proportion of purchases that are made by their new users. However, our analysis reveals that sequential recommendation models do not provide enough explore exposure to all items. Surprisingly, we find that some items receive zero explore exposure (i.e., these items will only be recommended to repeat users).

These findings matter because many sequential recommendation models suffer from the issue of *zero/less explore exposure*, which can influence long-term performance from the item perspective, i.e., it is unlikely to get such items exposed to new users.

Repetition "shortcuts" and inherent repetitive bias. Our consistent observation in the above analyses suggests that repeat-next users (that is, users who prefer to purchase a repeat item next) may act as a "shortcut" [11] to the optimization goal of sequential recommendation models, leading those model to recommend repetitive items even for explore-next users, i.e., repetitive bias. To investigate the potential impact of this shortcut on explore-next users, we design a counterfactual experiment: we remove all repeat-next users from the dataset and only train models based on explore-next users (that is, users who prefer to purchase a explore item next), so that there will be no shortcut during training and the model is optimized exclusively for the explore-next users, which can be seen as a pure exploration scenario.

We find that removing the shortcuts results in a higher degree of novelty of the recommendation (meaning that less repeat items are recommended to explore-next users). This confirms that the existence of shortcuts biases sequential recommendation (SR) models towards recommending repetitive items. Surprisingly, we also find that *sequential recommendation models will still recommend repeat items to users even in datasets with users who will only explore.* This means that SR models often fail to capture the simple characteristics of the pure exploration datasets and have an *inherent repetitive bias issue.*

Our analysis identifies the usage of *shared item embeddings* in the prediction layer as one potential cause of worsening the repetitive bias, as representations of user preferences inferred by the SR model tend to be highly similar to the item embeddings present in the input item sequence. We find that replacing shared item representations with independent item embeddings in the prediction layer alleviates this issue, thereby increasing the novelty of recommendations.

To complete our study and analysis of repetitive bias, we propose a remedy called the (3R) strategy, i.e., **r**emove **r**epeat items **r**ule, that simply removes repeat items from the predicted recommendation results. With this remedy, the accuracy of existing SR models in pure exploration scenario can be improved by a large margin.

Our findings matter because the issue of inherent repetitive bias impacts the performance of SR models in the pure exploration scenario. Future models should be evaluated more rigidly so as to determine where observed improvements come from.

Our contributions. The main contributions of this paper are:

- We analyze the accuracy of SR models through the lens of repeat and explore items. We confirm that the imbalance in performance and difficulty between the repetition task and exploration task known from other recommendation tasks also exists in the SR scenario. We point out evaluation issues of only using the overall average performance (in terms of accuracy) with a significance test.
- We generalize the perspective on repetition and exploration by adopting both a user-centered and an item-centered perspective. To the best of our knowledge, we are the first to propose *item explore exposure* and *item repeat exposure* to analyze the exposure allocation at a more fine-grained level.
- We demonstrate the importance of considering item explore exposure and show that several state-of-the-art SR models suffer from the problem of zero/few item explore exposure.
- We analyze the outcomes of our study by uncovering two key phenomena: (i) the impact of repetition "shortcuts": SR models may skew the recommendation towards repeat items by exploiting

shortcuts, which leads to a repetitive bias for explore-next users, and (ii) inherent repetitive bias. We investigate the differences between using shared item representations and independent item representations in the prediction layer and propose a remedy to eliminate the repetitive bias issue.

2 RELATED WORK OR BACKGROUND

In this section, we describe several empirical research lines in recommender systems that serve as the background of this work.

Sequential recommendation. Sequential item recommendation has been extensively studied. Several models employing deep learning techniques have been proposed [12, 13, 16, 19, 21, 30, 31, 37], such as RNN [12, 13], CNN [31], GNN [29, 37], contrastive learning [39], attention [19, 21], and self-attention [16, 30, 33, 41]. SASRec [16] was the first sequential recommendation model that employed a self-attention mechanism [33]. BERT4Rec [30] later upgraded the left-to-right training scheme in SASRec by using a bidirectional transformer with a Cloze task [32]. In addition, flexible orders [26], capturing repetition and exploration [28], and a consistent representation space [14] have all been found to improve the accuracy of the sequential recommendation.

Accuracy. There are several reproducibility and empirical studies focusing on accuracy-related metrics for recommender systems. Jannach and Ludewig [15] compare the performance of neural-based sequential models with nearest neighbor-based models; Petrov and Macdonald [25] evaluate the performance of BERT4Rec with different versions of implementations; Fang et al. [9] investigate several factors that influence the GRU4Rec performance; and, Zhao et al. [45] investigate the influence of different dataset splitting methods. However, these accuracy-oriented studies have primarily focused on the average performance. They do not provide a detailed assessment of performance for different user groups. More recently, Li et al. [20] have introduced a new evaluation perspective on the NBR task by differentiating between repetition (recommending items that users have purchased before) and exploration (recommending items that are new to the user) tasks in NBR. The authors highlight the difficulty of striking a balance between the two tasks. They analyze existing methods in NBR and conclude that the performance of many existing methods is mainly due to a (strong) bias towards the repetition task, at the expense of their ability to explore. Building on this study, the NBR models proposed in [2, 17] are designed to exploit those insights and improve their effectiveness.

However, the performance of sequential recommendation models w.r.t. repetition and exploration is still unexplored. This paper fills this gap by analyzing the impact of repetition and exploration in a sequential recommendation scenario.

Beyond accuracy. Apart from accuracy, diversity is another aspect to satisfy users' diversified demand [6, 27, 35, 43]. Recently, similar empirical and revisit studies [22, 40] have been performed to investigate the trade-off between accuracy and diversity. The notion of item exposure is used to measure item-side performance. It has become an important factor that models need to consider, as items and producers are important participants within a recommender system and the ecosystem in which it is deployed. Existing research w.r.t. item exposure is mostly focused on individual or group fairness, either on the customer-side, i.e., adopting a user-centered perspective [4], or on the provider-side, i.e., adopting an item-centered perspective [23, 42]; or two-sided [24, 36, 38].

Instead of analyzing the general exposure an item or group gets, as most prior work does, we are specifically interested in how sequential recommendation models allocate exposure in relation to repetition and exploration behavior.

3 PROBLEM FORMULATION AND DEFINITIONS

3.1 Sequential recommendation task

We use $\mathcal{I}$ and $\mathcal{U}$ for the sets of all items and users, respectively. Given a user $u \in \mathcal{U}$ and her historical item sequence $I_u = [i_1, i_2, \dots, i_t]$, where i_t denotes the item that the user interacted with at timestamp t, the sequential recommendation model $\mathcal{M}_{sq}$ infers the user's preferences from the historical sequence I_u and predicts the next items as recommendation results p_u^{t+1} at timestamp $t+1$. Formally:

$$p_u^{t+1} = \mathcal{M}_{sq}(I_u), \tag{1}$$

where p_u^{t+1} is a score distribution over the items. Usually, p_u^{t+1} is used to extract a ranked list of k items as the most probable items that u may interact with at timestamp $t+1$. Similarly to I_u, we use U_i to denote the sequence of users who have interacted with a given item $i \in \mathcal{I}$. When no confusion is possible, we use the same notations I_u and U_i for the set (instead of sequence) of historical items and users, respectively. In those cases, we define $\bar{I}_u = \mathcal{I} \setminus I_u$ and $\bar{U}_i = \mathcal{U} \setminus U_i$.

3.2 Repeat or explore

Using the historical interaction sequences I_u and U_i, for each user u and item i, we can divide both users and items as follows:

User-centered perspective. For a given user u, the items can be divided into *repeat items* I_u and *explore items* $\bar{I}_u$ (i.e., items that user u has not interacted with before).

Item-centered perspective. For a given item i, the users can be divided into *repeat users* U_i and *explore users* $\bar{U}_i$ (i.e., users who have not interacted with item i before).

Repeat-next user vs. explore-next user. Given the item the users will purchase next, the users can be divided into *repeat-next users* U_* (i.e., users who will purchase a repeat item in the next step) and *explore-next users* $\bar{U}_*$ (i.e., users who will purchase an explore item in the next step).

3.3 Explore exposure vs. repeat exposure

Item exposure. Conventional item exposure measures the number of times an item is recommended to users or the chance of an item being examined by the user. Besides, the position of an item in a recommendation list can influence its exposure, e.g., items at the top of the list are likely to receive more exposure than items at the bottom of the list. Usually, a click model C, which measures the likelihood that the user will examine the item in each position, is used in computing the item exposure, that is:

$$E_i@K = \sum_{u \in U_i} C_K(r_{u,i}), \tag{2}$$

where $r_{u,i}$ is the position of item i in the recommendation list shown to user u. In this study, we use the exposure model from the discounted cumulative gain (DCG) formula, i.e., $C_K(r_{u,i}) = (\mathbb{I}(r_{u,i} \leqslant K))/(\log_2(r_{u,i}+1))$.

The conventional definition of item exposure provided above does not account for the allocation of exposure to different types of users. From the repetition and exploration perspective, it is possible to evaluate the exposure allocation of the item to different types of users, i.e., *repeat users* and *explore users*. Formally, we propose item explore exposure and item repeat exposure as follows:

Item repeat exposure refers to the accumulated exposure that item i get from its repeat users U_i, that is:

$$RE_i@K = \sum_{u \in U_i} C_K(r_{u,i}). \tag{3}$$

Item explore exposure refers to the accumulated exposure that item i gets from its explore users $\bar{U}_i$, that is:

$$EE_i@K = \sum_{u \in \bar{U}_i} C_K(r_{u,i}). \tag{4}$$

Using the above two definitions, we define the Explore exposure ratio as the proportion of a given item's explore exposure from the total exposure it gets in the recommender system, that is:

$$EEr_i@K = \frac{EE_i@K}{E_i@K}. \tag{5}$$

This metric provides an individual-level assessment of exposure allocation. In the extreme case where $EEr_i@K = 0$, item i is only recommended to users who have purchased it before and will not be recommended to explore users.

4 EXPERIMENTAL SETUP

4.1 Research questions

In this study, we address the following main question:

(RQ1) How do the SR models perform w.r.t. repetition and exploration? Does the imbalance between repetition and exploration reported in prior work on NBR also exist in SR scenarios?

To analyze the answers to our main question, we pursue four additional, more specific questions:

(RQ2) Should we consider item explore exposure in the SR? How do the sequential recommendation models perform w.r.t. item explore exposure and item repeat exposure?

(RQ3) Does the repetition "shortcut" impose the SR models to recommend repeat items for explore-next users?

(RQ4) Does the repetitive bias of the sequential recommendation model still exist in a pure exploration scenario?

(RQ5) How can we avoid the potential effect of this repetitive bias?

4.2 Datasets

As our goal is to investigate the performance from the repetition and exploration perspective, we select two widely used sequential datasets with both repetition and exploration behavior:

Diginetica is a widely used dataset released in CIKM2016 Challenge, which includes user e-commerce search sessions with unique ids.[1]

Table 1: Statistics of the processed datasets. * RNU denotes repeat-next users; † ENU denotes explore-next users.

Dataset	#items	#users	# RNU*	# ENU†	ENU proportion
Diginetica	35,042	75,739	22,610	53,129	61.9%
Yoochoose	30,833	1,878,967	715,518	1,163,449	70.2%

Yoochoose is a widely used dataset released in the RecSys2015 Challenge, which contains a collection of sessions from a retailer, and each session in the dataset represents a series of click events performed by a user during the session.[2]

We follow the widely used preprocessing procedure in previous works, i.e, "5-core". Specifically, we remove items that are purchased/viewed less than 5 times and remove users whose interaction sequence length is less than 5. We set the maximum length of a sequence to 50 and any sequences longer than 50 are truncated. We split each dataset into train, validation, and test partitions using a leave-one-out strategy: for each item sequence, we hold the final interaction for the test set, the second last interaction for the validation set, and the third last interaction for the train set. The statistics of the pre-processed datasets are shown in Table 1.

4.3 Methods

Methods selection. The purpose of this study is to provide insights w.r.t. performance evaluation and model design from a novel angle, rather than to track and confirm the best or latest sequential recommendation model. Thus, we consider the following aspects to select the methods we want to analyze:

- **Influential**: the selected method should be highly-cited and influential, which continue serving as competitive baselines in sequential recommendation research.
- **Representative**: the selected methods should have diverse representation techniques, which continue serving as the backbone of various sequential recommendation models.
- **Consistency**: the selected methods should follow the same paradigm of modeling, which only takes users' historical item sequence as input to generate the users' preference representation.[3]

Methods. Following the criteria listed above, we select several highly-cited methods with representative techniques (i.e., RNN, CNN, GNN, transformers, BERT) as follows:

- **GRU4Rec** is a representative method that uses a recurrent neural network (i.e. a GRU) to model users' sequential behavior [13].
- **Caser** is a representative method that uses a CNN to model users' sequential behavior [31].
- **SRGNN** is a representative method that uses a graph neural network (GNN) to model user historical sequence [37].
- **SASRec** is a representative method that employs a left-to-right Transformer model to capture users' sequential behavior [16].
- **BERT4Rec** is a representative method that employs a bi-directional transformer model and introduces the Cloze task to train the model [18].
- **RepeatNet** is a representative method that models the users' preference w.r.t. repetition and exploration, and uses separate decoders for repeat item and explore item prediction [28].

[1] https://competitions.codalab.org/competitions/11161

[2] https://www.kaggle.com/datasets/chadgostopp/recsys-challenge-2015

[3] Note that we do not include sequential recommendation methods with a user embedding or any additional information as input to maintain fairness and consistency.

Table 2: The contribution of repeat-next users to the average overall performance w.r.t. Recall@1.

Dataset	GRU4Rec	Caser	SRGNN	BERT4Rec	SASRec	RepeatNet
Diginetica	71.5%	81.1%	88.0%	78.5%	86.5%	100%
Yoochoose	90.5%	89.2%	94.3%	92.9%	94.8%	100%

Configurations. For the neural-based sequential recommendation methods listed above, we use the implementations in the Recbole open-source project and then integrate them into our pipeline. We follow the hyper-parameter settings suggested in Recbole. The embedding size is tuned on $\{32, 64, 128\}$ for all methods based on the validation set to achieve their best performance. For BERT4Rec and SASRec, we use two stacked transformer layers with 8 heads. For all methods, the dropout ratio is set to 0.1 and the Adam optimizer is employed with a learning rate of 0.001.

All the training is performed using TITAN X GPUs with 12G memory. We repeat our experiments 5 times and report the average performance. We share both our dataset processing scripts, the source code, and the hyper-parameters we use in an anonymous repository.[4]

4.4 Metrics

We use three widely used metrics for the sequential recommendation problem, i.e., Recall@K, MRR@K, and NDCG@K, to measure accuracy. In the sequential recommendation task, *Recall* measures the ability to find a relevant item that meets the user's preference; *NDCG* and *MRR* are metrics that also consider the order of the relevant items. For these three accuracy-oriented metrics, the higher the value, the better the performance.

We also use Novelty$_u$@K to measure the novelty of the recommendation, that is:

$$\text{Novelty}_u@K = \frac{\sum_{r=1}^{K} h(u,r) \cdot \log_2(r+1)}{\sum_{r=1}^{K} \log_2(r+1)} \tag{6}$$

where $h(u,r) = 1$ if the r^{th} item in the recommended list to user u is a explore item, otherwise $h(u,r) = 0$. Explore-next users prefer higher novelty, while repeat-next users prefer lower novelty. We will later describe our proposed metrics in later sections to remain focused. In this paper, we consider the metrics with $K \in \{1, 3\}$, as a higher K will lead to a passive increase w.r.t. the novelty and the item explore exposure we will discuss below.[5]

5 REPETITION ACCURACY AND EXPLORATION ACCURACY

Evaluation. In this section, we aim to gain an understanding of accuracy from the repetition and exploration angle and find potential issues w.r.t. only using average overall accuracy. Specifically, apart from the average overall accuracy, we also examine a more fine-grained level, analyzing the accuracy (Recall, NDCG, MRR) and novelty (Novelty) w.r.t. two user groups, i.e., repeat-next users U_* and explore-next users $\bar{U}_*$.

Table 3: The novelty of the recommendation for repeat-next users and explore-next users.

	Diginetica				Yoochoose			
	Novelty@1		Novelty@3		Novelty@1		Novelty@3	
Method	RNU	ENU	RNU	ENU	RNU	ENU	RNU	ENU
GRU4Rec	0.567	0.675	0.725	0.779	0.223	0.362	0.488	0.577
Caser	0.237	0.304	0.634	0.660	0.265	0.436	0.496	0.597
SRGNN	0.145	0.208	0.447	0.479	0.099	0.172	0.408	0.476
SASRec	0.231	0.285	0.511	0.556	0.085	0.134	0.423	0.492
BERT4Rec	0.385	0.483	0.530	0.599	0.110	0.183	0.409	0.479
RepeatNet	0	0	0.010	0.032	0	0	0.108	0.150

Results. The experimental results w.r.t. different accuracy metrics are shown in Figure 1 and Table 2. We have the following observations: (i) there is a large imbalance in recommendation accuracy between repeat-next users and explore-next users, where all models achieve noticeably higher recommendation accuracy (across different metrics) w.r.t. repeat-next users than explore-next users; (ii) compared to explore-next users, repeat-next users account for a relatively small proportion of the user population, whereas they contribute to a large proportion of the average performance; (iii) the absolute difference in recommendation accuracy between different methods w.r.t. explore-next users is smaller than the difference w.r.t. repeat-next users; and (iv) a higher average overall accuracy does not necessarily link to the improvement w.r.t. the recommendation accuracy across all users, e.g., RepeatNet achieves the best overall accuracy in most cases on Diginetica, whereas it has the lowest accuracy on both datasets w.r.t. explore-next users.

The above results answer RQ1 and confirm that the findings w.r.t. the imbalance between repetition and exploration in NBR setting generalize to the sequential item recommendation scenario.

In general, the expected novelty for repeat-next users is 0, meaning that only repeat items are recommended, while the expected novelty for explore-next users is 1, indicating that only explore items are recommended. The experimental results w.r.t. the novelty over different types of users are shown in Table 3. We have the following observations: (i) different methods exhibit diverse performance in terms of the novelty of the recommendation; for instance, GRU4Rec has relatively high novelty, while RepeatNet and SRGNN have much lower novelty compared to GRU4Rec; and (ii) the novelty of recommendations to explore-next users is slightly higher than for repeat-next users in most cases, indicating that these sequential recommendation models have the ability to identify user preferences towards repetition and exploration to some degree.

Sacrificing the performance for specific users. When there is a huge imbalance between the recommendation performance w.r.t. different groups of user, using the average overall performance to represent the performance of a method has a risk of hiding and sacrificing the performance for users for whom the recommendation task relatively (more) difficult (e.g., users who prefer to explore, in this paper). For instance, compared to BERT4Rec and SASRec, RepeatNet and SRGNN can achieve higher overall performance by sacrificing the performance for a large proportion of users who prefer to explore on the Diginetica dataset.

[4] https://github.com/liming-7/Repetition-exploration-SR

[5] For example, if the length of a user's historical sequence is 5, there are at most 5 different items that could be regarded as repeat items, so the recommendation list with a size of 10 will always contain at least 5 explore items.

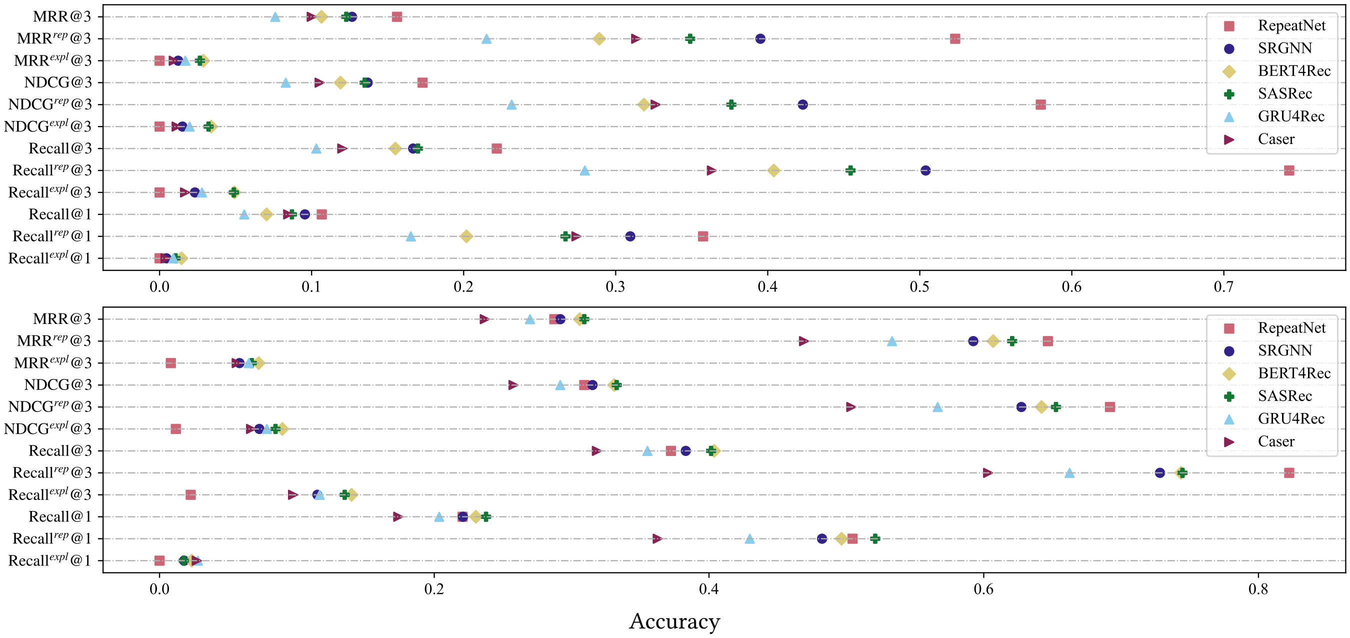

Figure 1: The recommendation accuracy for all users, repeat-next users and explore-next users. Top: Diginetica; bottom: Yoochoose.

Significance testing. In recommender system research, we often use a significance test when comparing the performance of models. If the p-value is less than a pre-determined level of significance (usually 0.05 or 0.01), we usually claim something like "The proposed model A significantly outperforms baseline B". Based on the findings above, we want to caution against over-reliance on the successful outcomes of a significance test in the context of SR. Our concern stems from our experimental results that show that RepeatNet achieves higher accuracy scores than SASRec, with a paired significance test p-value below 0.05. However, SASRec performs better for users who prefer to explore, who account for over 60% of the users.

Lessons. The findings concerning reproducibility that we have listed above, confirm that the analysis of repetition and exploration is also important, but neglected, in SR scenarios, just as in NBR scenarios, which motivates us to perform a deeper analysis of SR models from several angles w.r.t. repetition and exploration.

Furthermore, upon drilling down, we have found that, when comparing sequential recommendation models on a dataset containing users who prefer to repeat, we should be aware that the widely-used average overall accuracy with a significance test may not fully represent the models' recommendation accuracy for all user groups. Instead, we should also: (i) evaluate the accuracy for repeat-next users and explore-next users separately and perform separate significance tests on these two averages; and (ii) check the actual accuracy distribution to know whether the method favors a specific user group over another.

6 EXPLORE EXPOSURE AND REPEAT EXPOSURE

In this section, we first illustrate the importance of analyzing item explore exposure and then perform an analysis of this property.

6.1 Importance of item explore exposure

In order to demonstrate the importance of analyzing item explore exposure and answer RQ2, we perform the following two analyses:

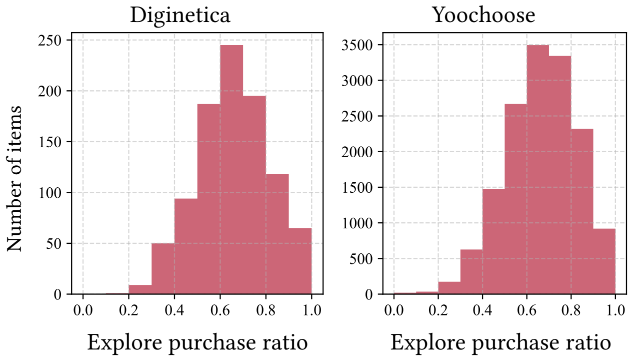

Figure 2: Distribution of items across different exploratory purchase ratios.

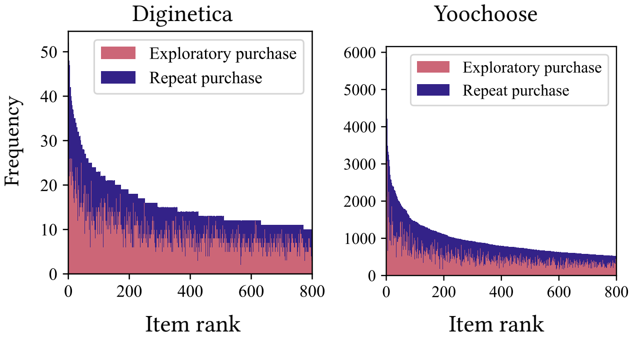

Figure 3: Distribution of purchases across different items.

Table 4: Proportion of items with zero explore exposure; analyzed on top-500 popular next items with $K = 1$.

Dataset	GRU4Rec	Caser	SRGNN	BERT4Rec	SASRec	RepeatNet
Diginetica	11.8%	16.2%	38.2%	19.4%	35.7%	100%
Yoochoose	6.7%	2.8%	16.7%	18.7%	33.9%	99.7%

Exploratory purchase. The number of exploratory purchases for an item reflects the expected number of times that the item will be purchased by a new user u_i^{novel} in the future. A large number of exploratory purchases of an item indicates that it should be recommended to a large number of explore users, i.e., require explore exposure, otherwise, it will lose a large potential user purchase and never be known by these potential explore users. We rank the items based on their total purchases and find a substantial number of exploratory purchases across different items in both datasets, as shown in Figure 3.

Exploratory purchase ratio. The exploratory purchase ratio (*EPr*) of an item refers to the proportion of exploratory purchases within all future purchases of this item. For statistical analysis of the exploratory purchase ratio, we ignore items with less than 10 future purchases for the sake of confidence. From Figure 2, we observe the distribution is right-skewed, i.e., the *EPr* is more spread out towards a higher value, which indicates a large proportion of the future purchases are made by explore users. An item with a high *EPr* should be recommended more to potential explore users than repeat users (a.k.a the item should get more explore exposure than repeat exposure). An extreme case with $EPr_i = 1$ indicates that all future purchases will be made by explore users for the item i, so it is meaningless for giving repeat exposure to this item.

6.2 Less/zero explore exposure issue

Evaluation. The exploratory purchase ratio (*EPr*) provides an expected exposure distribution between explore exposure and repeat exposure, which can be seen as the expected explore exposure ratio. Therefore, we can evaluate whether the SR models provide items with enough explore exposure compared to what is expected by computing the difference between the exploratory purchase ratio (*EPr*) and the explore exposure ratio (*EEr*), that is:

$$\Delta_i^{\mathrm{E}}@K = EPr_i@K - EEr_i@K.$$

Specifically, we first rank items according to their future total purchases and then calculate the average Δ_i^{E} of items with top-Q total purchases and non-zero total exposure for the following reasons: (i) items in the catalog are not equally important, and fewer total purchases of item indicate that only a small number of users will prefer this item, and (ii) we focus on the item exposure distribution, and analyzing the repeat exposure and explore exposure distribution of items with zero exposure is meaningless.

Findings. To answer RQ2, we re-evaluate the performance of SR models w.r.t. their exposure allocation to items. The results of our analysis are shown in Figure 4. For all methods on both datasets, we observe that the average Δ^{E} is negative, which indicates that items tend to receive less explore exposure than expected, and their exposure is biased towards repeat exposure.

Note that RepeatNet has the lowest Δ^{E} score, which can be seen as the lower-bound of Δ^{E}.[6] Moreover, SRGNN and SASRec are very close to this lower-bound w.r.t. $\Delta^{\mathrm{E}}@1$ on both datasets. To further investigate the potential issues w.r.t. SR models, we analyze the proportion of items that will only be recommended to repeat users (i.e., zero explore exposure) in the sequential next-item recommendation scenario (i.e. $K = 1$). The results are shown in Table 4. Surprisingly, we find that a non-negligible proportion of items suffer from the zero explore exposure issue, which is a severe problem as these items may never be discovered or seen by their potential new users.

Lessons. According to the findings in this section, we should be aware that items need explore exposure and that SR models suffer from a less/zero explore exposure issue w.r.t. a neglected proportion of items. Instead of only analyzing the overall item exposure from an item-centered perspective, we should also analyze the exposure distribution w.r.t. repetition and exploration.

7 REPETITIVE BIAS

Pure exploration. An important task of recommender systems is to connect users with items that they have never seen; there is a large proportion of explore-next users, who would like to explore items. From the analysis in Section 5, the imbalance in difficulty between the repetition task and the exploration task suggests that the existence of repeat-next users is a "shortcut" to the optimization goal of accuracy-oriented SR models, leading those models to recommend repetitive items even for explore-next users.

Specifically, we remove all repeat samples (i.e., repeat-next users) from the train, validation, and test set to ensure there will be no shortcut during training and optimization, so that the model will be specifically trained and optimized for explore-next users. Note that this constructed subset can be regarded as a pure exploration scenario, as all the training, validation, and test ground-truth labels in this constructed subset are explore items.

Influence of removing repetition shortcuts. From Figure 5, we observe that removing shortcuts leads to a higher novelty for all methods on both datasets. This illustrates that the presence of repeat-next users makes the model more likely to recommend repeat items even for explore-next users, this answers RQ3.

A counterintuitive finding. In this pure exploration scenario, we surprisingly find that *some sequential recommendation models will still recommend repeat items to users even in datasets with pure exploration.* This finding is counterintuitive since humans can easily notice the basic characteristics of this scenario, i.e., there are no repeat items in the ground-truth labels in the training, validation, and test set. Whereas, many complex models fail to detect this simple pattern of the dataset and have an inherent repetitive bias issue, which is a serious pitfall that results in poor user experience.[7]

Shared embeddings vs. independent embeddings. We suspect that the user's preference representation inferred by the SR model

[6] Even equipped with a module to identify whether a user prefers to explore or repeat, RepeatNet can only recommend repeat items to the user (Novelty = 0), which also means the item will only be recommended to users who purchased them before.

[7] RepeatNet can avoid the inherent repetitive issue since it has an indicator to identify repetition and exploration. BERT4Rec employs a self-supervised training objective, so it is not that surprising to have repetitive bias.

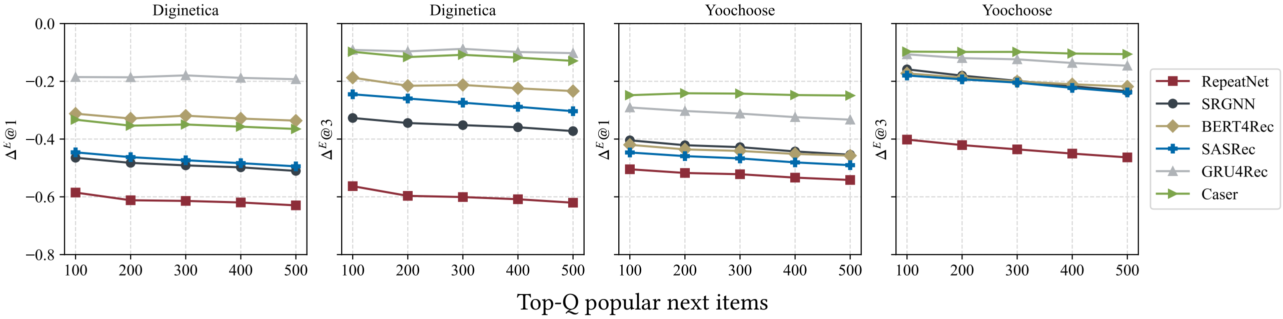

Figure 4: Difference between expected EEr and actual EEr of the over different item groups.

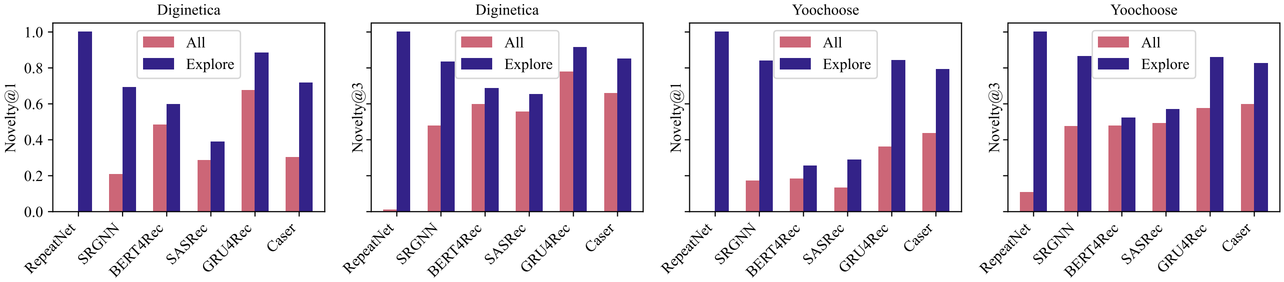

Figure 5: The recommendation novelty for explore-preferred users: train using all vs. using pure exploration.

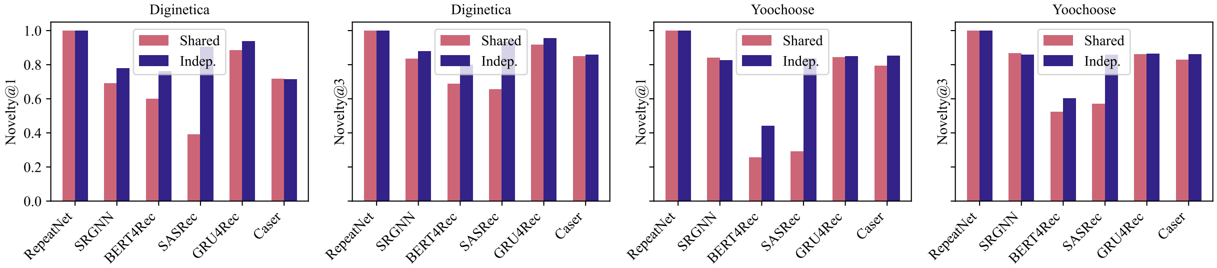

Figure 6: The recommendation novelty w.r.t. explore-next users: shared embedding vs. independent embedding.

will be similar to the item embeddings within the model's input sequence. Therefore, repeat items will be ranked high in the recommendation list when using the dot product between user representation and item embedding as the prediction layer. The prediction layer typically shares item embeddings with the input layer, which reduces model size and may reduce overfitting [16, 30, 37]. To understand whether using shared embeddings in the prediction layer will exacerbate bias towards repetitive items, we replace the shared item representation with an independent item representation to conduct another group of experiments.

From the results in Figure 6, we find that replacing shared item embeddings with independent item embeddings in the prediction layer can indeed alleviate this repetitive issue and increase the novelty of recommendations for explore-next users. The results confirm that using shared embeddings in the prediction layer contributes to the issue of repetitive bias of SR models, which answers RQ4.

A simple remedy: the 3R strategy. However, we can also see that using independent embeddings does not entirely address the repetitive bias issue of SR models. A straightforward method to eliminate the repetitive bias in the pure exploration scenario is to post-process recommendation results according to the scenario's characteristics. We propose a remedy called the 3R strategy, i.e., **r**emove **r**epeat item from **r**ule, which simply removes the repeat items in the recommendation in the pure exploration scenario. From the experiment results in Table 5, we observe that: (i) simply adopting the 3R strategy can easily bring substantial improvements across all methods, and (ii) shared embeddings with the 3R strategy outperforms independent embeddings with the 3R strategy in terms of recommendation accuracy for all methods on both datasets. 3R strategy is the answer to RQ5 in the pure exploration scenario.

Lessons. According to the analysis above, when comparing SR models in a pure exploration scenario, we should be aware that: (i) many complex SR models have the inherent repetitive bias issue, which negatively impacts their performance in a pure exploration SR scenario, (ii) the SR models may perform differently when the input item embeddings are not shared with the prediction layer, and using shared embedding may exacerbate the inherent repetitive bias issue, and (iii) a higher recommendation accuracy can be easily achieved by addressing the repetitive bias using the 3R strategy to post-process recommendations for a pure exploration scenario.

Additionally, it is important for future models to be rigorously evaluated to check where improvements truly come from.

8 CONCLUSION

In this paper, we have investigated and revisited sequential recommendation from the repetition and exploration perspective. Taking

Table 5: The recommendation accuracy w.r.t. explore-next users of SR models with 3R strategy. S and I denote using shared and independent embeddings, respectively. We exclude RepeatNet here since it does not have repetitive bias.

		Diginetica-Expl.		Yoochoose-Expl.	
Method	Mode	Recall@1	NDCG@3	Recall@1	NDCG@3
SRGNN	S	0.0107	0.0197	0.0716	0.1155
	S+3R	0.0135	0.0232	0.0823	0.1302
	I	0.0104	0.0180	0.0640	0.1061
	I+3R	0.0125	0.0203	0.0754	0.1212
BERT4Rec	S	0.0140	0.0309	0.0270	0.0886
	S+3R	0.0259	0.0493	0.1102	0.1767
	I	0.0158	0.0323	0.0399	0.0942
	I+3R	0.0230	0.0431	0.1021	0.1643
SASRec	S	0.0119	0.0323	0.0330	0.0947
	S+3R	0.0300	0.0520	0.1035	0.1681
	I	0.0234	0.0377	0.0824	0.1356
	I+3R	0.0256	0.0403	0.0991	0.1562
GRU4Rec	S	0.0091	0.0161	0.0614	0.1021
	S+3R	0.0106	0.0183	0.0725	0.1172
	I	0.0066	0.0118	0.0577	0.0961
	I+3R	0.0073	0.0130	0.0699	0.1123
Caser	S	0.0053	0.0108	0.0415	0.0740
	S+3R	0.0072	0.0131	0.0595	0.0980
	I	0.0042	0.0087	0.0480	0.0817
	I+3R	0.0058	0.0107	0.0548	0.0935

lessons learned in a NBR scenario as our starting point, we analyzed several representative SR models in multiple ways: (i) from a user-centered perspective, where we analyze the accuracy and novelty w.r.t. repeat-next users and explore-next users, and (ii) from an item-centered perspective, where we define explore exposure and repeat exposure to measure exposure allocation. We have also investigated the repetitive bias of SR models w.r.t. the recommendation for explore-next users from the following aspects: (i) the repetition "shortcuts," and (ii) shared embedding and independent embedding.

Findings. We arrive at several important findings and discover some issues w.r.t. SR models: (i) as in NBR, in SR too there is a huge imbalance between repetition and exploration, and SR models perform much better for repeat-next users than explore-next users; (ii) a higher average performance can be achieved by sacrificing the performance for a large proportion of users, which indicates that our widely used evaluation strategy, i.e., "overall performance with significance test", hides important details about the effectiveness of SR models; (iii) many SR models suffer from a less/zero explore exposure issue, i.e., items are mostly (or even only) recommended to their repeat users; (iv) the existence of repetition "shortcuts" increases the repetitive bias w.r.t. the recommendation for explore-next users; (v) many SR models suffer from an inherent repetitive bias (i.e., they still recommend repeat items even in the pure exploration scenario), and using shared embeddings will exacerbate this inherent repetitive bias; and (vi) a simple strategy for post-processing the recommendations of SR models may lead to substantial improvements in a pure-exploration scenario.

Broader implications. Our work highlights the following important lessons that practitioners and researchers should follow: (i) in a SR scenario with both repetition and exploration, instead of only relying on the average overall accuracy with a significance test, we should also evaluate the performance of repeat-next users and explore-next users separately, and check the distribution of performance results across users; (ii) in a pure exploration SR scenario, we should be aware of the inherent repetitive bias issue, and use the 3R strategy to post-process SR models when using them as baselines; and (iii) on a two-sided platform, SR practitioners should also check the explore exposure and the exposure allocation of items to ensure that items will not only get exposed to their repeat users and have explore exposure to reach potential exploring consumers.

Our analyses show that using the repetition "shortcut" in SR scenarios with repetition behavior and addressing the repetitive bias in SR scenarios with pure exploration may lead to substantial improvements w.r.t. recommendation accuracy of SR models. Given that many recent SR models were evaluated without separately considering repetition and exploration performance, it is unclear whether the improvements observed come from improving the model overall or from leveraging shortcuts that improve repetition at the expense of exploration. For evaluations conducted in an exploration scenario, it is unclear which improvements would remain after mitigating the models' repetitive bias with the 3R strategy.

Limitations and future work. Our analyses mainly focus on the neural-based SR models that have been published in recent years, ignoring classic machine learning-based and neighbor-based methods. Another limitation is that we only focus on repetition and exploration, but there might be other factors that also lead to a performance imbalance in the SR scenario. We focus on analyzing the exposure distribution of items and uncovered the limited item explore exposure issue; it would be interesting to consider how to avoid this issue.

ACKNOWLEDGMENTS

We would like to thank our reviewers for their constructive feedback. This research was partially supported by Ahold Delhaize, the China Scholarship Council (grant #20190607154), Elsevier and the Netherlands Organisation for Scientific Research (NWO) under project nr 652.002.001, and the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Ashton Anderson, Ravi Kumar, Andrew Tomkins, and Sergei Vassilvitskii. 2014. The Dynamics of Repeat Consumption. In *Proceedings of the 23rd International Conference on World Wide Web*. 419–430.

[2] Mozhdeh Ariannezhad, Sami Jullien, Ming Li, Min Fang, Sebastian Schelter, and Maarten de Rijke. 2022. 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*. 1240–1250.

[3] Austin R Benson, Ravi Kumar, and Andrew Tomkins. 2016. Modeling User Consumption Sequences. In *Proceedings of the 25th International Conference on World Wide Web*. 519–529.

[4] Jesús Bobadilla, Raúl Lara-Cabrera, Ángel González-Prieto, and Fernando Ortega. 2020. Deepfair: Deep Learning for Improving Fairness in Recommender Systems. *arXiv preprint arXiv:2006.05255* (2020).

[5] Jun Chen, Chaokun Wang, Jianmin Wang, and S Yu Philip. 2016. Recommendation for Repeat Consumption from User Implicit Feedback. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (2016), 3083–3097.

[6] Wanyu Chen, Pengjie Ren, Fei Cai, Fei Sun, and Maarten de Rijke. 2020. Improving End-to-end Sequential Recommendations with Intent-aware Diversification. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 175–184.

[7] Zhilong Chen, Hancheng Cao, Huangdong Wang, Fengli Xu, Vassilis Kostakos, and Yong Li. 2020. Will You Come back/Check-in Again? Understanding Characteristics Leading to Urban Revisitation and Re-check-in. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 4. 1–27.

[8] Michael D. Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and Discrimination in Retrieval and Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1403–1404.

[9] Hui Fang, Danning Zhang, Yiheng Shu, and Guibing Guo. 2020. Deep Learning for Sequential Recommendation: Algorithms, Influential Factors, and Evaluations. *ACM Transactions on Information Systems* 39, 1 (2020), Article 10.

[10] Yingqiang Ge, Shuchang Liu, Ruoyuan Gao, Yikun Xian, Yunqi Li, Xiangyu Zhao, Changhua Pei, Fei Sun, Junfeng Ge, Wenwu Ou, and Yongfeng Zhang. 2021. Towards Long-Term Fairness in Recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 445–453.

[11] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut Learning in Deep Neural Networks. *Nature Machine Intelligence* 2, 11 (2020), 665–673.

[12] Balázs Hidasi and Alexandros Karatzoglou. 2018. Recurrent Neural Networks with Top-k Gains for Session-based Recommendations. In *Proceedings of the 27th ACM International Conference on Information & Knowledge Management*. 843–852.

[13] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based Recommendations with Recurrent Neural Networks. *arXiv preprint arXiv:1511.06939* (2015).

[14] Yupeng Hou, Binbin Hu, Zhiqiang Zhang, and Wayne Xin Zhao. 2022. Core: Simple and Effective Session-based Recommendation within Consistent Representation Space. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*. 1796–1801.

[15] Dietmar Jannach and Malte Ludewig. 2017. When Recurrent Neural Networks Meet the Neighborhood for Session-based Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. 306–310.

[16] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive Sequential Recommendation. In *2018 IEEE International Conference on Data Mining*. 197–206.

[17] Ori Katz, Oren Barkan, Noam Koenigstein, and Nir Zabari. 2022. Learning to Ride a Buy-Cycle: A Hyper-Convolutional Model for Next Basket Repurchase Recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 316–326.

[18] Duc-Trong Le, Hady W Lauw, and Yuan Fang. 2019. Correlation-sensitive Next-basket Recommendation. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 2808–2814.

[19] Jing Li, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Tao Lian, and Jun Ma. 2017. Neural Attentive Session-based Recommendation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1419–1428.

[20] Ming Li, Sami Jullien, Mozhdeh Ariannezhad, and Maarten de Rijke. 2023. A Next Basket Recommendation Reality Check. *ACM Transactions on Information Systems* 41, 4 (2023), Article 116.

[21] Qiao Liu, Yifu Zeng, Refuoe Mokhosi, and Haibin Zhang. 2018. STAMP: Short-term Attention/Memory Priority Model for Session-based Recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1831–1839.

[22] Malte Ludewig and Dietmar Jannach. 2018. Evaluation of Session-based Recommendation Algorithms. *User Modeling and User-Adapted Interaction* 28 (2018), 331–390.

[23] Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 429–438.

[24] Mohammadmehdi Naghiaei, Hossein A. Rahmani, and Yashar Deldjoo. 2022. 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*. 770–779.

[25] Aleksandr Petrov and Craig Macdonald. 2022. A Systematic Review and Replicability Study of BERT4Rec for Sequential Recommendation. In *Proceedings of the 16th ACM Conference on Recommender Systems*. 436–447.

[26] Mingda Qian, Xiaoyan Gu, Lingyang Chu, Feifei Dai, Haihui Fan, and Bo Li. 2022. Flexible Order Aware Sequential Recommendation. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*. 109–117.

[27] Massimo Quadrana, Paolo Cremonesi, and Dietmar Jannach. 2018. Sequence-Aware Recommender Systems. *Comput. Surveys* 51, 4 (2018), 1–36.

[28] Pengjie Ren, Zhumin Chen, Jing Li, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019. RepeatNet: A Repeat Aware Neural Recommendation Machine for Session-based Recommendation. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. 4806–4813.

[29] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* 20, 1 (2008), 61–80.

[30] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. 2019. BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 1441–1450.

[31] Jiaxi Tang and Ke Wang. 2018. Personalized Top-n Sequential Recommendation via Convolutional Sequence Embedding. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*. 565–573.

[32] Wilson L Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly* 30, 4 (1953), 415–433.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *arXiv preprint arXiv:1706.03762* (2017).

[34] Shoujin Wang, Liang Hu, Yan Wang, Longbing Cao, Quan Z Sheng, and Mehmet Orgun. 2019. Sequential Recommender Systems: Challenges, Progress and Prospects. *arXiv preprint arXiv:2001.04830* (2019).

[35] Shoujin Wang, Liang Hu, Yan Wang, Quan Z Sheng, Mehmet Orgun, and Longbing Cao. 2019. Modeling Multi-purpose Sessions for Next-item Recommendations via Mixture-channel Purpose Routing Networks. In *International Joint Conference on Artificial Intelligence*. 3771–3777.

[36] Haolun Wu, Bhaskar Mitra, Chen Ma, Fernando Diaz, and Xue Liu. 2022. Joint Multisided Exposure Fairness for Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 703–714.

[37] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. 2019. Session-based Recommendation with Graph Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 346–353.

[38] Yao Wu, Jian Cao, Guandong Xu, and Yudong Tan. 2021. TFROM: A Two-Sided Fairness-Aware Recommendation Model for Both Customers and Providers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1013–1022.

[39] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive Learning for Sequential Recommendation. In *2022 IEEE 38th International Conference on Data Engineering*. 1259–1273.

[40] Qing Yin, Hui Fang, Zhu Sun, and Yew-Soon Ong. 2022. Understanding Diversity in Session-Based Recommendation. *arXiv preprint arXiv:2208.13453* (2022).

[41] Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Jundong Li, and Zi Huang. 2022. Self-Supervised Learning for Recommender Systems: A Survey. *arXiv preprint arXiv:2203.15876* (2022).

[42] Meike Zehlike and Carlos Castillo. 2020. Reducing Disparate Exposure in Ranking: A Learning To Rank Approach. In *Proceedings of The Web Conference 2020*. 2849–2855.

[43] Mi Zhang and Neil Hurley. 2008. Avoiding Monotony: Improving the Diversity of Recommendation Lists. In *Proceedings of the 2008 ACM conference on Recommender systems*. 123–130.

[44] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *Comput. Surveys* 52, 1 (2019), 1–38.

[45] Wayne Xin Zhao, Junhua Chen, Pengfei Wang, Qi Gu, and Ji-Rong Wen. 2020. Revisiting Alternative Experimental Settings for Evaluating Top-n Item Recommendation Algorithms. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2329–2332.