Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning

Dušan Stamenković*
University of Novi Sad, Serbia
dusan.stamenkovic@dmi.uns.ac.rs

Alexandros Karatzoglou
Google Research, United Kingdom
alexkz@google.com

Ioannis Arapakis
Telefonica Research, Spain
ioannis.arapakis@telefonica.com

Xin Xin
Shandong University, China
xinxin@sdu.edu.cn

Kleomenis Katevas
Telefonica Research, Spain
kleomenis.katevas@telefonica.com

摘要

Since the inception of Recommender Systems (RS), the accuracy of the recommendations in terms of relevance has been the golden criterion for evaluating the quality of RS algorithms. However, by focusing on item relevance, one pays a significant price in terms of other important metrics: users get stuck in a "filter bubble" and their array of options is significantly reduced, hence degrading the quality of the user experience and leading to churn. Recommendation, and in particular session-based/sequential recommendation, is a complex task with multiple - and often *conflicting* objectives - that existing state-of-the-art approaches fail to address.

In this work, we take on the aforementioned challenge and introduce *Scalarized Multi-Objective Reinforcement Learning (SMORL)* for the RS setting, a novel Reinforcement Learning (RL) framework that can effectively address multi-objective recommendation tasks. The proposed SMORL agent augments standard recommendation models with additional RL layers that enforce it to simultaneously satisfy three principal objectives: *accuracy*, *diversity*, and *novelty* of recommendations. We integrate this framework with four state-of-the-art session-based

*Full affiliation is Faculty of Sciences, University of Novi Sad. This work is done when taking an internship in Telefonica Research, Spain, and it was supported by C4IIOT.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning

Dušan Stamenković*
University of Novi Sad, Serbia
dusan.stamenkovic@dmi.uns.ac.rs

Alexandros Karatzoglou
Google Research, United Kingdom
alexkz@google.com

Ioannis Arapakis
Telefonica Research, Spain
ioannis.arapakis@telefonica.com

Xin Xin
Shandong University, China
xinxin@sdu.edu.cn

Kleomenis Katevas
Telefonica Research, Spain
kleomenis.katevas@telefonica.com

摘要

自推荐系统（Recommender Systems, RS）诞生以来，推荐的相关性准确性一直是评估RS算法质量的黄金标准。然而，专注于物品相关性往往会在其他重要指标上付出重大代价：用户被困在“过滤气泡”中，选择范围大幅度减少，从而降低用户体验质量并导致用户流失。推荐，尤其是基于会话/顺序的推荐，是一个复杂的任务，具有多个且往往是冲突的目标，而现有的最先进方法未能解决这些问题。

在这项工作中，我们挑战上述问题，引入标量化多目标强化学习（*Scalarized Multi-Objective Reinforcement Learning, SMORL*）用于RS设置，这是一种新颖的强化学习（Reinforcement Learning, RL）框架，可以有效解决多目标推荐任务。所提出的SMORL智能体（agent）增强了标准推荐模型，采用额外的RL层，使其同时满足三个主要目标：准确性、多样性和新颖性。我们将该框架与四个最先进的基于会话的推荐模型进行整合，并与一个仅专注于准确性的单一目标RL智能体进行比较。我们在两个真实世界数据集上的实验结果显示，

*Full affiliation is Faculty of Sciences, University of Novi Sad. This work is done when taking an internship in Telefonica Research, Spain, and it was supported by C4IIOT.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Woodstock '18, June 03–05, 2018, Woodstock, NY
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

recommendation models and compare it with a single-objective RL agent that only focuses on accuracy. Our experimental results on two real-world datasets reveal a substantial increase in aggregate diversity, a moderate increase in accuracy, reduced repetitiveness of recommendations, and demonstrate the importance of reinforcing diversity and novelty as complementary objectives.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Retrieval models and ranking**; • **Diversity and novelty in information retrieval**;

KEYWORDS

Recommendation; Reinforcement Learning; Multi-Objective Reinforcement Learning; Diversity; Novelty

ACM Reference Format:

Dušan Stamenković, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. 2018. Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Whether in the context of entertainment, social networking or e-commerce, the sheer number of choices that modern Web users face nowadays can be overwhelming. Contrary to the common belief that more options are always better, selections made from large assortments can lead to a choice overload [17] and impair users' capacity for rational decision making. Simply put, when presented with large array situations (e.g., limitless products to purchase from or media content to consume), users are at higher risk of feeling like they made the wrong decision and experience regret, which can degrade the quality of experience with an online service or platform. The problem becomes further aggravated when one is inclined to consider the costs and benefits of all alternative options.

Recommender Systems (RS) alleviate this *paradox of choice* [32] by acting as *second-order strategies* [35] that facilitate access to relevant information and improve the browsing experience [16, 43]. Hence, in settings where the abundance of options can result in unsatisfying choices or, even worst, abandonment, the user experience is ultimately determined by the RS capacity to filter irrelevant content and recommend only items regarded as desirable. So far, the main focus of the research community in the area of RS has been placed on designing algorithms that can identify and recommend relevant content. However, while doing so, they tend to optimise (for the most part) mainstream metrics such as accuracy, at the expense of other content-derived qualitative aspects. In this work, the term "accuracy" denotes the performance of the RS in terms of ranking relevant items in the offline test set, and it should not be mistaken for accuracy in classification tasks.

In recent years, diversity and novelty of recommendations have been recognized as important factors for promoting user engagement, since recommending a diverse set of relevant items is more likely to satisfy users' variable needs. For example, Hu and Pu [15] report a strong positive correlation between diversity of recommendations and ease of use, perceived usefulness, and intentions to use the system. Therefore, a RS that suggests strictly relevant items to a user who just purchased an espresso machine will, most likely, end up recommending more coffee machines, while the preferred set of recommendations would include coffee mugs, cleaning equipment, coffee beans, so to speak. In the former case, users will get to interact only with a small subspace of the available item space [28] and, according to the "law of diminishing marginal returns", the utility of the recommendations

聚合多样性有显著增加，准确性适度提高，推荐的重复性降低，并证明了增强多样性和新颖性作为互补目标的重要性。

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Retrieval models and ranking**; • **Diversity and novelty in information retrieval**;

KEYWORDS

Recommendation; Reinforcement Learning; Multi-Objective Reinforcement Learning; Diversity; Novelty

ACM Reference Format:

Dušan Stamenković, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas. 2018. Choosing the Best of Both Worlds: Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

无论是在娱乐、社交网络还是电子商务的背景下，现代网络用户面临的选择数量往往令人不知所措。与普遍认为的“更多的选项总是更好”的看法相反，从大量选项中所做的选择可能导致选择过载 [17]，并削弱用户理性决策的能力。简单来说，当用户面临大量选择的情况（例如，购买无尽的产品或消费媒体内容）时，他们更有可能感觉自己做出了错误的决定并经历后悔，这会降低他们与在线服务或平台的体验质量。当用户倾向于考虑所有替代选项的成本和收益时，这一问题变得更加严重。

推荐系统（RS）通过作为*第二阶策略* [35]，缓解这一*选择悖论* [32]，促进对相关信息的访问并改善浏览体验 [16, 43]。因此，在选项丰富可能导致不满意选择或更糟糕的放弃的情况下，用户体验最终取决于推荐系统过滤不相关内容并仅推荐被视为理想项的能力。迄今为止，推荐系统领域的研究重点主要放在设计能够识别和推荐相关内容的算法上。然而，在这样做的过程中，他们往往过度优化（大部分情况下）主流指标，如准确性，牺牲其他内容导出的定性方面。在这项工作中，术语“准确性”表示推荐系统在离线测试集中的相关项排名表现，不应与分类任务中的准确性混淆。

近年来，推荐的多样性和新颖性被认为是促进用户参与的重要因素，因为推荐多样的相关项更有可能满足用户的变动需求。例如， Hu and Pu [15] 报告了推荐多样性与可用性、感知有用性和使用意图之间的强正相关。因此，向刚刚购买了浓缩咖啡机的用户推荐严格相关的物品，最终很可能会推荐更多的咖啡机，而理想的推荐集合应包括咖啡杯、清洁设备、咖啡豆，等等。在前一种情况下，用户只会与可用物品空间的一个小子空间进行互动 [28]，根据“边际效益递减法则”，随着用户一次又一次地接触到类似内容，推荐的效用最终将下降。

基于会话的推荐被引入作为推荐系统的另一种与行业相关的方法。在基于会话的推荐中，一个顺序模型（例如，RNN [14] 或 transformer [19, 37]）以自我监督的方式进行训练，以预测序列中下一个项，而不是一些“外部”标签 [14, 19, 43]。该训练过程受到语言建模任务的启发，其中，给定一个词序列输入，语言模型预测下一个最可能出现的词 [24]。然而，这种训练方法也可能产生次优推荐，因为损失函数完全由模型预测与序列中的实际项之间的不匹配定义。在这种损失函数下训练的模型仅关注匹配用户可能生成的点击序列，而放弃其他期望的目标。例如，服务提供商可能希望促进那些将汇聚到购买、提高用户满意度、多样化用户与物品互动并促进长期参与的推荐。然而，为了使推荐系统优化以达到上述目标，需要用可微分函数捕捉这些目标，而这并不是一件简单的事情。因此，在重要目标只能以非可微分函数/指标的形式展示的领域，多目标优化（MOO）的使用受到很大限制。

多样性和推荐项目列表的新颖性与销售多样性的增加相关[9]，并通过推荐所谓“长尾”中的不太流行项目来解决“赢家通吃”的问题。来自多样化推荐列表的一个项目更有可能是新颖的，即用户通常不会与

will eventually degrade as users are exposed to similar content, over and over again.

Session-based recommendation has been introduced as an alternative, industry-relevant approach to RS. In session-based recommendation, a sequential model (e.g., a RNN [14] or a transformer [19, 37]) is trained in a self-supervised fashion to predict the next item in the sequence itself, instead of some "external" labels [14, 19, 43]. This training process was inspired by language modelling tasks where, given a word sequence input, the language model predicts the most likely word to appear next [24]. However, this training method can also produce sub-optimal recommendations, since the loss function is defined purely by the mismatch between model predictions and the actual items in the sequence. Models trained under such a loss function focus only on matching the sequence of clicks a user may generate, while forfeiting other desirable objectives. For example, a service provider may want to promote recommendations that will converge to purchases, increase user satisfaction, diversify user-item interactions and promote long-term engagement. Nevertheless, in order to optimize an RS towards said objectives, one needs to capture them with a differentiable function, which is not a trivial task. Therefore, the use of multi-objective optimization (MOO) is heavily limited in areas where important objectives can only be presented in a form of non-differentiable functions/metrics.

Diversity and novelty of recommended item lists are correlated with increased diversity of sales [9], and address the "winner-takes-all" problem by recommending less popular items from the so-called "long-tail'. An item from a diverse recommendation list is more likely to be novel, i.e., an item that the user would not normally interact with. This is supported by prior work, which suggests that most users appreciate novel and less popular recommendations [21, 44]. Recommendation models trained with simple supervised learning may encounter difficulties in addressing the above recommendation expectations and the multi-objective nature of many online tasks.

To address the current challenges, we expand on the idea of utilising RL in the RS setting and introduce a *Scalarized Multi-Objective RL* (SMORL) approach. SMORL uses a single RL agent to simultaneously satisfy three, potentially conflicting, objectives: i) promote clicks, ii) diversify the set of recommendations, and iii) introduce novel items, while at the same time optimising for relevance. The model focuses on the chosen rewards while maintaining high relevance ranking performance. More specifically, given a generative sequential or session-based recommendation model, the (final) hidden state of the model can be seen as it' s output layer, since it is multiplied with the last (dense softmax) layer to generate the recommendations [14, 19, 43]. We augment these models with multiple final output layers. The conventional self-supervised head, is trained with the cross-entropy loss to perform ranking, while the SMORL part is simultaneously trained to modify the rankings of the self-supervised head. The RL heads can be seen as regularizers that introduce more diverse and novel recommendations, while the ranking-based supervised head can provide more robust learning signals (including negative signals) for parameter updates. One of the main advantages of using MORL instead of MOO in the context of RS is the possibility of using non-differentiable functions for reward system that the RL agent uses to regularize the base model.

Previous attempts of balancing accuracy with diversity and novelty included re-ranking of the final set of recommendations or training of multiple models and the use of genetic algorithms to aggregate those models [31], whereas our approach relies on training a single model and using the SMORL framework to balance the principal recommendation objectives. We argue that this framework can be easily extrapolated to other domains such as music, video, and news recommendations (by using

之互动的项目。这得到了以往工作的支持，研究表明大多数用户欣赏新颖且不太流行的推荐 [21, 44]。使用简单的监督学习训练的推荐模型可能在满足上述推荐期望和许多在线任务的多目标性质方面遇到困难。

为了解决当前的挑战，我们扩展了在推荐系统（RS）环境中利用强化学习（RL）的想法，并引入了一种标量化多目标强化学习（Scalarized Multi-Objective RL, SMORL）方法。SMORL使用单个RL智能体（agent）同时满足三个潜在冲突的目标：i) 促进点击，ii) 多样化推荐集，以及 iii) 引入新颖项目，同时优化相关性。该模型专注于所选择的奖励（reward），同时保持高的相关性排名性能。更具体地说，给定一个生成式的序列或基于会话的推荐模型，模型的（最终）隐藏状态可以视为它的输出层，因为它与最后的（密集softmax）层相乘以生成推荐 [14, 19, 43]。我们为这些模型增添了多个最终输出层。传统的自监督头通过交叉熵损失进行训练以执行排名，而SMORL部分则同时训练以修改自监督头的排名。RL头可以视为正则化器，用于引入更多样化和新颖的推荐，而基于排名的监督头可以为参数更新提供更强的学习信号（包括负信号）。在推荐系统的背景下，使用多目标强化学习（MORL）而不是多目标优化（MOO）的主要优势之一是能够使用不可微分的函数作为RL智能体用于正则化基础模型的奖励系统。

之前平衡准确性与多样性和新颖性的尝试包括对最终推荐集的重新排名，或训练多个模型并使用遗传算法聚合这些模型 [31]，而我们的方法依赖于训练单个模型，并使用SMORL框架来平衡主要推荐目标。我们认为，这个框架可以很容易地推广到其他领域，如音乐、视频和新闻推荐（通过使用嵌入系统 [3, 23]），在这些领域中，多样性和新颖性是高价值的指标。总之，我们的工作做出了以下贡献：

- 我们设计了一种新颖的多样性奖励（diversity reward），该奖励利用了物品嵌入空间（item embedding space）。
- 我们设计了一种用于评估推荐系统（RS）的新指标，该指标衡量推荐的重复性。
- 据我们所知，我们首次在推荐系统（RS）的背景下应用多目标强化学习（MORL），并探索这一方法所提供的众多可能性和未来研究方向。
- 我们引入了SMORL，它驱动自监督的推荐系统 (RS) 产生更准确、多样化和新颖的推荐。我们将四个最先进的推荐模型 (model) 集成到所提出的框架中。
- 我们在两个真实的电子商务数据集上进行实验，展示了更少重复的推荐集，在聚合多样性指标上有显著改善（高达20%），同时保持，甚至提升了所有四个状态（state）-of-the-art（最先进）模型的准确性。

2 RELATED WORK

多个基于深度学习的方法已被提出，用于有效地建模用户交互序列以适应推荐系统（RS）。Hidasi et al. [14] 使用门控循环单元（GRU） [8] 来建模用户会话，而 Tang and Wang [36] 和 Yuan et al. [43] 则使用卷积神经网络（CNN）来捕捉序列信号。Kang and McAuley [19] 在序列推荐领域利用了著名的 Transformer [37]，取得了良好的结果。所有这些模型都可以作为基础模型，其输入是用户-项目交互的序列，输出是描述相应用户状态的潜在表征（representation）。

也有多个尝试将强化学习（RL）应用于推荐系统。在 off-policy 设置中，Chen et al. [5] 和 Zhao et al. [45] 提出了使用倾向分数进行 off-policy 校正，但由于高方差导致训练困难。有模型（model-based）RL 方法 [6, 34, 47] 首先建立一个模型以模拟环境，从而避免任何与 off-policy 训练相关的问题。然而，这两阶段方法在很大程度上依赖于模拟器的准确性。Xin et al. [42] 引入了 SQN 和 SAC，这两个自我监督（self-supervised）RL框架为推荐系统增强了推荐模型，增加了两个输出层（heads）。第一个头是基于交叉熵监督损失，而另一个 RL 头则基于双重 Q 学习（Double Q-learning）[11]。尽管 SQN 和 SAC 提高了性能，但它们仅通过促进用户可能进行的点击和购买来增加准确性。然而，准确的推荐系统不一定是有效的：真正的价值在于提

embedding systems [3, 23]), where diversity and novelty are high-value metrics. In summary, our work makes the following contributions:

- We devise a novel diversity reward that utilises the item embedding space.
- We devise a novel metric for evaluation of RS that measures repetitiveness of recommendations.
- To the best of our knowledge, we apply Multi-Objective Reinforcement Learning (MORL) in the setting of RS for the first time and explore some of the many possibilities and future research directions that this approach offers.
- We introduce SMORL that drives the self-supervised RS to produce more accurate, diverse and novel recommendations. We integrate four state-of-the-art recommendation models into the proposed framework.
- We conduct experiments on two real-world e-commerce datasets and demonstrate less repetitive recommendations sets, significant improvements in aggregate diversity metrics (up to 20%), all while maintaining, or even improving accuracy for all four state-of-the-art models.

2 RELATED WORK

Several deep learning-based approaches that model the user interaction sequences effectively have been proposed for RS. Hidasi et al. [14] used gated recurrent units (GRU) [8] to model user sessions, while Tang and Wang [36] and Yuan et al. [43] used convolutional neural networks (CNN) to capture sequential signals. Kang and McAuley [19] exploited the well-known Transformer [37] in the field of sequential recommendation, with promising results. All of these models can serve as the base model whose input is a sequence of user-item interactions and the output is a latent representation that describes the corresponding user state.

Several attempts to use RL for RS have also been made. In the off-policy setting, Chen et al. [5] and Zhao et al. [45] proposed the use of propensity scores to perform off-policy correction, but with training difficulties due to high-variance. Model-based RL approaches [6, 34, 47] first build a model to simulate the environment, in order to avoid any issues with off-policy training. However, these two-stage approaches depend heavily on the accuracy of the simulator. Xin et al. [42] introduced SQN and SAC, two self-supervised RL frameworks for RS that augment the recommendation model with two output layers (heads). First head is based on the cross-entropy supervised loss, while the other RL head is based on the Double Q-learning [11]. Although SQN and SAC improve performance, they only increase accuracy by promoting clicks and purchases that a user might make. However, an accurate RS is not necessarily a useful one: real value lies in suggesting items that users would likely not discover for themselves, that is, in the novelty and diversity of recommendations [12]. Improving accuracy typically decreases diversity and novelty, which can occur when RL is deployed to regularize session-based RS (see discussion in Section 5). A decrease of aggregate diversity can impact the user experience and satisfaction with the RS [15]. Anderson et al. [1] also report that current recommendations discourage diverse user-item interactions.

Diversifying recommendations and introducing novel recommendations were recently recognized as important factors for improving RS. Early efforts focused on post-processing methods that aimed to balance accuracy and diversity [2, 29, 33]. In order to mitigate issues with significant cumulative loss on the ranking function, personalized ranking methods were proposed [7]. Chen et al. [4] tried to address the issues of post-processing methods that consider only pairwise measures of diversity and ignore correlations between items, by proposing the probabilistic model Determinantal Point Process [22] that captures the correlation between items using a kernel matrix. Once

示用户可能不会自行发现的项目，即推荐的新颖性和多样性 [12]。提高准确性通常会降低多样性和新颖性，这可能在 RL 被应用于常规化基于会话的推荐系统时出现（见第 5 节中的讨论）。聚合多样性的减少会影响用户体验和对推荐系统的满意度 [15]。Anderson et al. [1] 也报告称，当前的推荐会抑制多样化的用户-项目交互。

最近，多样化推荐和引入新颖推荐被认为是改善推荐系统的重要因素。早期的努力集中在后处理方法上，旨在平衡准确性和多样性 [2, 29, 33]。为了缓解在排名函数上显著累积损失的问题，提出了个性化排名方法 [7]。Chen et al. [4] 尝试解决只考虑对偶测量多样性而忽视项目间相关性的后处理方法的问题，提出了概率模型——行列式点过程（Determinantal Point Process）[22]，该模型使用核矩阵捕捉项目间的相关性。一旦学习了这个矩阵，许多抽样技术便可生成多样的项目集合 [4, 38, 40]。这些模型在最佳情况下实现了准确性和多样性之间的权衡。另一方面，SMORL显著增加了多样性，并略微提高了准确性。

在 RL 设置中，Zheng et al. [46] 关注于探索-利用策略，以促进多样性，通过随机选择当前推荐项目邻域中的随机项目候选。 Hansen et al. [10] 提出了一个基于 RL 的抽样排名器，生成多样项目的排名列表。该模型是一个简单的排名器，模型本身并没有学习生成多样项目集合，而学习过程利用了 REINFORCE 算法 [41]，该算法已知存在高方差问题。最后，以往在推荐系统中优化多个目标的尝试依赖于帕累托优化（Pareto-Optimization），使用网格搜索 [31] 或多梯度下降 [25]。然而，按定义，一个帕累托最优解并不一定比其他帕累托最优解在所有目标上更优。

3 MULTI-OBJECTIVE RL FOR RS

让 $\mathcal{I}$ 表示整个物品集，那么用户-物品交互序列可以表示为 $\mathbf{x}_{1:t} = \{x_1, x_2, ..., x_{t-1}, x_t\}$，其中 $x_i \in \mathcal{I}\,(0 < i \leq t)$ 是在时间戳 i 上交互的[1] 物品的索引。下一个物品推荐的目标是向用户推荐最适合其当前兴趣的物品 $\mathbf{x}_{t+1}$，给定之前交互序列 $\mathbf{x}_{1:t}$。

从多目标强化学习（MORL）的角度来看，下一个物品推荐任务可以被表述为一个多目标马尔可夫决策过程（MOMDP）[39]，其中推荐智能体与环境 $\mathcal{E}$（用户）通过顺序推荐物品来最大化折扣累计奖励。MOMDP 可以通过元组 $(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R}, \rho_0, \gamma)$ 定义，其中：

- $\mathcal{S}$: 一个连续状态空间，描述用户状态。用户在时间戳 t 的状态可以定义为 $\mathbf{s}_t = G(\mathbf{x}_{1:t}) \in \mathcal{S}(t > 0)$，其中 G 是一个将在 Section 4 中讨论的顺序模型 (sequential model)。
- $\mathcal{A}$: 一个包含候选项的离散动作空间。智能体（agent）的动作 a 是推荐所选项。在离线强化学习（offline RL）设置中，我们要么从用户-项目交互中提取时间戳 t 的动作，即 $a_t = x_{t+1}$，要么通过将其设置为从自监督层获得的最佳预测来进行设置。状态-动作对 $(\mathbf{s}_t, a_t)$ 的"好坏"由其多目标 Q 价值函数 $\mathbf{Q}(\mathbf{s}_t, a_t)$ 来描述。
- $\mathbf{P} : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ 是状态转移概率 $p(\mathbf{s}_{t+1}|\mathbf{s}_t, a_t)$，即当智能体 (agent) 采取动作 (action) a_t 时，从 $\mathbf{s}_t$ 到 $\mathbf{s}_{t+1}$ 的状态 (state) 转移概率。
- $\mathbf{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^m$ 是向量值奖励函数[2]，其中 $\mathbf{r}(\mathbf{s}, a)$ 表示在状态 $\mathbf{s}$ 下采取动作 a 所获得的即时奖励。
- ρ_0 是初始状态 (state) 分布，其中 $\mathbf{s}_0 \sim \rho_0$.
- $\gamma \in [0, 1]$ 是未来奖励的折扣因子。对于 $\gamma = 0$，智能体（agent）只考虑即时奖励，而对于 $\gamma = 1$，所有未来奖励都被完全考虑，除了当前动作（action）的奖励。

目标是寻找MORL（多目标强化学习）智能体（agent）在MOMDP（多目标马尔可夫决策过程）中以目标策略 $\pi_\theta(a|\mathbf{s})$ 的形式解决问题，使得根据 $\pi_\theta(a|\mathbf{s})$ 采样的轨迹能够实现最大期望累积奖励（reward）：

$$\max_{\pi_\theta} \mathbb{E}_{\tau\sim\pi_\theta}\left[f\left(\mathbf{R}(\tau)\right)\right], \text{ where } \mathbf{R}(\tau) = \sum_{t=0}^{|\tau|} \gamma^t \mathbf{r}(\mathbf{s}_t, a_t)$$

[1]在现实世界场景中,可能有不同种类的交互。例如,在电子商务中,交互可以是点击、购买、购物车添加等。在音乐推荐中,交互可以通过歌曲的播放时间、歌曲的听歌次数等来表征。

[2]每个分量对应一个目标。

this matrix is learned, many sampling techniques can generate a diverse set of items [4, 38, 40]. These models achieve a trade-off between accuracy and diversity at best. On the other hand, SMORL significantly increases diversity and slightly improves the accuracy.

In the RL setting, Zheng et al. [46] focused on exploration-exploitation strategies for promoting diversity, by randomly choosing random item candidates in the neighborhood of the current recommended item. Hansen et al. [10] proposed a RL sampling-based ranker that produces a ranked list of diverse items. This model is a simple ranker and the model itself doesn't learn to produce diverse set of items, while the learning process utilizes the REINFORCE algorithm [41] which is known to suffer from high-variance. Finally, prior attempts to optimize multiple objectives in the setting of RS relied on Pareto-Optimization using grid search [31] or multi-gradient descent [25]. However, by definition, one Pareto optimal solutions is not necessarily better than other Pareto optimal solution with respect to all objectives.

3 MULTI-OBJECTIVE RL FOR RS

Let $\mathcal{I}$ denote the whole item set, then a user-item interaction sequence can be represented as $\mathbf{x}_{1:t} = \{x_1, x_2, ..., x_{t-1}, x_t\}$, where $x_i \in \mathcal{I}\,(0 < i \leq t)$ is the index of the interacted[1] item at timestamp i. The goal of next item recommendation is recommending the item $\mathbf{x}_{t+1}$ to users that will best suit their current interests, given the sequence of previous interactions $\mathbf{x}_{1:t}$.

From the perspective of MORL, the next item recommendation task can be formulated as a Multi-Objective Markov Decision Process (MOMDP) [39], in which the recommendation agent interacts with the environments $\mathcal{E}$ (users) by sequentially recommending items to maximize the discounted cumulative rewards. The MOMDP can be defined by tuples of$(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{R}, \rho_0, \gamma)$ where:

- $\mathcal{S}$: a continuous state space that describes the user state. The state of the user at timestamp t can be defined as $\mathbf{s}_t = G(\mathbf{x}_{1:t}) \in \mathcal{S}\,(t > 0)$, where G is a sequential model that will be discussed in Section 4.
- $\mathcal{A}$: a discrete action space that contains candidate items. The action a of the agent is to recommend the selected item. In the offline RL setting, we either extract the action at timestamp t from the user-item interaction, i.e., $a_t = x_{t+1}$, or by setting it to a top prediction obtained from the self-supervised layer. The "goodness" of a state-action pair $(\mathbf{s}_t, a_t)$ is described by its multi-objective Q-value function $\mathbf{Q}(\mathbf{s}_t, a_t)$.
- $\mathbf{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the state transition probability $p(\mathbf{s}_{t+1}|\mathbf{s}_t, a_t)$, i.e., a probability of state transition from $\mathbf{s}_t$ to $\mathbf{s}_{t+1}$ when agent takes action a_t.
- $\mathbf{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^m$ is the vector-valued reward function[2], where $\mathbf{r}(\mathbf{s}, a)$ denotes the immediate reward by taking action a at state $\mathbf{s}$.
- ρ_0 is the initial state distribution with $\mathbf{s}_0 \sim \rho_0$.
- $\gamma \in [0, 1]$ is the discount factor for future rewards. For $\gamma = 0$, the agent only considers the immediate reward, while for $\gamma = 1$, all future rewards are regarded fully except the one of the current action.

The goal of the MORL agent is to find a solution to a MOMDP in a form of target policy $\pi_\theta(a|\mathbf{s})$ so that sampling trajectories according to $\pi_\theta(a|\mathbf{s})$ would lead to the maximum expected cumulative reward:

$$\max_{\pi_\theta} \mathbb{E}_{\tau \sim \pi_\theta}\left[f\big(\mathbf{R}(\tau)\big)\right], \text{ where } \mathbf{R}(\tau) = \sum_{t=0}^{|\tau|} \gamma^t \mathbf{r}(\mathbf{s}_t, a_t)$$

[1]In a real world scenario there may be different kinds of interactions. For instance, in e-commerce, the interactions can be clicks, purchases, basket additions, and so on. In music recommendation, the interactions can be characterized by the play time of a song, the number of times a song was listened, etc.

[2]Each component corresponds to one objective.

Choosing the Best of Both Worlds:
Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning Woodstock '18, June 03–05, 2018, Woodstock, NY

其中 $f : \mathbb{R}^m \mapsto \mathbb{R}$ 是一个标量化函数，而 $\theta \in \mathbb{R}^d$ 表示策略 (policy) 参数。期望是在轨迹 $\tau = (\mathbf{s}_0, a_0, \mathbf{s}_1, a_1 ...)$ 上取的，该轨迹是通过根据目标策略执行动作 (action) 获得的：$\mathbf{s}_0 \sim \rho_0, a_t \sim \pi_\theta(\cdot|\mathbf{s}_t), \mathbf{s}_{t+1} \sim \mathbf{P}(\cdot|\mathbf{s}_t, a_t)$。

标量化函数 f 将多目标 Q 价值 (Q-values) $\mathbf{Q}(\mathbf{s}_t, a_t)$ 和奖励 (reward) 函数 $\mathbf{r}(\mathbf{s}_t, a_t)$ 映射到一个标量值，即用户效用。在本文中，我们专注于线性 f；每个目标 i 被赋予一个重要性，即权重 (weight) $w_i, i = 1, ..., m$，使得标量化函数变为 $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$，其中 $\mathbf{w} = [w_1, .., w_m]$。

4 MODEL AND TRAINING

我们将下一项推荐任务视为一个（自监督）多类分类问题，并构建一个顺序模型，该模型接收用户-项目交互序列 $\mathbf{x}_{1:t} = [x_1, x_2, ..., x_{t-1}, x_t]$ 作为输入，并生成 n 个分类对数 $\mathbf{y}_{t+1} = [y_1, y_2, ..., y_n] \in \mathbb{R}^n$，其中 n 是候选项目的数量。然后，我们可以从 y_{t+1} 中选择 top-k 项作为我们在时间戳 $t + 1$ 的推荐列表。每个候选项目对应于一个类别。

通常可以使用生成序列模型 $G(\cdot)$ 将输入序列映射到隐藏状态 $\mathbf{s}_t = G(\mathbf{x}_{1:t})$。这作为一个通用编码器函数。基于获得的隐藏状态，我们可以利用一个简单的解码器将隐藏状态映射到分类对数 $y_{t+1} = d(\mathbf{s}_t)$。可以将解码器函数 d 定义为一个简单的全连接层或与候选项目嵌入的内积 [14, 19, 43]。在本工作中，我们使用全连接层。最后，我们通过优化分类对数 y_{t+1} 基于交叉熵损失 L_s 来训练我们的推荐模型。交叉熵损失的优化将促进正对数的高价值，而用户未互动的项目将被"惩罚"，从而产生强烈的负学习信号。这个负信号对于基础模型的学习至关重要，因为 SMORL头仅对正动作提供强梯度，即 top-1 项。此外，由于顺序推荐模型 G 已经将输入序列编码为潜在表征 $\mathbf{s}_t$，我们直接使用 $\mathbf{s}_t$ 作为 RL 头的当前状态，而无需引入单独的 RL 模型。我们堆叠额外的全连接层来计算位于 G 之上的一维 Q-值：

$$Q_z(\mathbf{s}_t, a_t) = \delta(\mathbf{s}_t \mathbf{H}_z + b_z) = \delta(G(\mathbf{x}_{1:t})\mathbf{H}_z + b_z)$$

其中 $z \in \{\text{acc}, \text{div}, \text{nov}\}$，$\delta$ 表示激活函数，而 $\mathbf{H}_z$ 和 b_z 是 Q-learning 输出层的可学习参数。SMORL 部分随后将计算出的准确性（accuracy）、多样性（diversity）和新颖性（novelty） Q-值堆叠成一个向量值 Q-值函数：

$$\mathbf{Q}(\mathbf{s}_t, a_t) = \left[Q_{\text{acc}}(\mathbf{s}_t, a_t), Q_{\text{div}}(\mathbf{s}_t, a_t), Q_{\text{nov}}(\mathbf{s}_t, a_t)\right] \quad (1)$$

为了学习向量值 Q 函数并解决 MORL 任务，标量化深度 Q 学习 (SDQL) [27] 扩展了流行的 DQN 算法 [26]，通过引入一个标量化函数 f。在每个时间步 t，Q 网络在从经验池 (buffer) D 中获得的经验元组 $(\mathbf{s}_t, a_t, r_t, s_{t+1})$ 的小批量上优化损失 L_{SDQL}：

$$\begin{aligned} L_{\text{SDQL}} &= (f(\mathbf{y}_t^{\text{SDQL}}(\mathbf{s}_t, a_t) - \gamma \mathbf{Q}(\mathbf{s}_{t+1}, a_t)))^2 \\ &= (\mathbf{w}^\top(\mathbf{y}_t^{\text{SDQL}}(\mathbf{s}_t, a_t) - \gamma \mathbf{Q}(\mathbf{s}_{t+1}, a_t)))^2 \end{aligned} \quad (2)$$

在这里，$\mathbf{y}_t^{\text{SDQL}}(\mathbf{s}_t, a_t) = \mathbf{r}_t + \gamma \mathbf{Q}'(\mathbf{s}_{t+1}, \text{argmax}_{a'}[\mathbf{w}^\top \mathbf{Q}'(\mathbf{s}_{t+1}, a')])$，其中$\mathbf{Q}'$是目标网络。向固定目标网络进行训练防止了状态间近似误差的快速传播，而采样经验进行训练（经验回放）则提高了样本效率，并减少了训练样本间的相关性。

在生成推荐时，我们仍然从监督头中返回前k个项目。SMORL头作为基础推荐模型G的正则化器，通过评估推荐的顶部项目质量，根据预定义的奖励（reward）设置和标量化函数f，即目标的重要性，对其进行微调。

4.1 Reinforcing Accuracy

为了使基础模型 G 学会提供更相关的推荐，我们扩展了 [42] 并将准确性奖励定义为

$$r_{\text{acc}}(\mathbf{s}_t, a_t) = r_{\text{acc}}(a_t) = 1, \quad a_t \text{ is a clicked item} \quad (3)$$

根据奖励的定义，当模型（model）匹配序列中的下一个点击项目时，将会获得奖励。Xin 等人 [42] 建议同时对点击和购买使用奖励。然而，在本工作中，我们提出了一种可以很容易从电子商务推广到其他相

where $f : \mathbb{R}^m \mapsto \mathbb{R}$ is a scalarization function, while $\theta \in \mathbb{R}^d$ denotes the policy parameters. The expectation is taken over trajectories $\tau = (\mathbf{s}_0, a_0, \mathbf{s}_1, a_1 ...)$, obtained by performing actions according to the target policy: $\mathbf{s}_0 \sim \rho_0, a_t \sim \pi_\theta(\cdot|\mathbf{s}_t), \mathbf{s}_{t+1} \sim \mathbf{P}(\cdot|\mathbf{s}_t, a_t)$.

A scalarization function f maps the multi-objective Q-values $\mathbf{Q}(\mathbf{s}_t, a_t)$ and a reward function $\mathbf{r}(\mathbf{s}_t, a_t)$ to a scalar value, i.e., the user utility. In this paper, we focus on linear f; each objective i is given an importance, i.e. weight $w_i, i = 1, ..., m$ such that the scalarization function becomes $f_\mathbf{w}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, where $\mathbf{w} = [w_1, .., w_m]$.

4 MODEL AND TRAINING

We cast the task of next item recommendation as a (self-supervised) multi-class classification problem and build a sequential model that receives user-item interaction sequence $\mathbf{x}_{1:t} = [x_1, x_2, ..., x_{t-1}, x_t]$ as an input and generates n classification logits $\mathbf{y}_{t+1} = [y_1, y_2, ..., y_n] \in \mathbb{R}^n$,where n is the number of candidate items. We can then choose the top-k items from y_{t+1} as our recommendation list for timestamp $t + 1$. Each candidate item corresponds to a class.

Typically one can use a generative sequence model $G(\cdot)$ to map the input sequence into a hidden state $\mathbf{s}_t = G(\mathbf{x}_{1:t})$. This serves as a general encoder function. Based on the obtained hidden state, we can utilize a simple decoder to map the hidden state to the classification logits as $y_{t+1} = d(\mathbf{s}_t)$. One can define the decoder function d as a simple fully connected layer or the inner product with candidate item embeddings [14, 19, 43]. In this work, we make use of the fully connected layer. Finally, we train our recommendation model by optimizing the cross-entropy loss L_s based on the logits y_{t+1}. Optimization of the cross-entropy loss will push the positive logits to high values, while the items that user did not interact with will be "penalised", which will result in a strong negative learning signal. This negative signal is essential for learning in the base model, since the SMORL head provides strong gradients only for positive actions, i.e., top-1 items. Furthermore, due to the fact that the sequential recommendation model G has already encoded the input sequence into a latent representation $\mathbf{s}_t$, we directly use $\mathbf{s}_t$ as the current state for the RL head without the need to introduce a separate RL model. We stack additional fully connected layers to calculate one-dimensional Q-values on top of G:

$$Q_z(\mathbf{s}_t, a_t) = \delta(\mathbf{s}_t \mathbf{H}_z + b_z) = \delta(G(\mathbf{x}_{1:t})\mathbf{H}_z + b_z)$$

where $z \in \{\text{acc}, \text{div}, \text{nov}\}$, δ denotes the activation function, while $\mathbf{H}_z$ and b_z are learnable parameters of the Q-learning output layer. The SMORL part then stacks computed accuracy, diversity, and novelty Q-values into a vector-valued Q-value function:

$$\mathbf{Q}(\mathbf{s}_t, a_t) = \left[Q_{\text{acc}}(\mathbf{s}_t, a_t), Q_{\text{div}}(\mathbf{s}_t, a_t), Q_{\text{nov}}(\mathbf{s}_t, a_t)\right] \quad (1)$$

In order to learn vector-valued Q-functions and tackle MORL tasks, *Scalarized Deep Q-learning* (SDQL) [27] extends the popular DQN algorithm [26], by introducing a scalarization function f. At every time step t, Q-network is optimized on the loss L_{SDQL} computed on a mini-batch of experience tuples $(\mathbf{s}_t, a_t, r_t, s_{t+1})$ obtained from experience buffer D:

$$\begin{aligned} L_{\text{SDQL}} &= (f(\mathbf{y}_t^{\text{SDQL}}(\mathbf{s}_t, a_t) - \gamma \mathbf{Q}(\mathbf{s}_{t+1}, a_t)))^2 \\ &= (\mathbf{w}^\top(\mathbf{y}_t^{\text{SDQL}}(\mathbf{s}_t, a_t) - \gamma \mathbf{Q}(\mathbf{s}_{t+1}, a_t)))^2 \end{aligned} \quad (2)$$

where $\mathbf{y}_t^{\text{SDQL}}(\mathbf{s}_t, a_t) = \mathbf{r}_t + \gamma \mathbf{Q}'(\mathbf{s}_{t+1}, \text{argmax}_{a'}[\mathbf{w}^\top \mathbf{Q}'(\mathbf{s}_{t+1}, a')])$, and $\mathbf{Q}'$ being the target network. Training towards a fixed target network prevents approximation errors from propagating too quickly from state to state, and sampling experiences to train on (experience replay) increases sample efficiency and reduces correlation between training samples.

When generating recommendations, we still return the top-k items from the supervised head. The SMORL head acts as a regularizer of the base recommendation

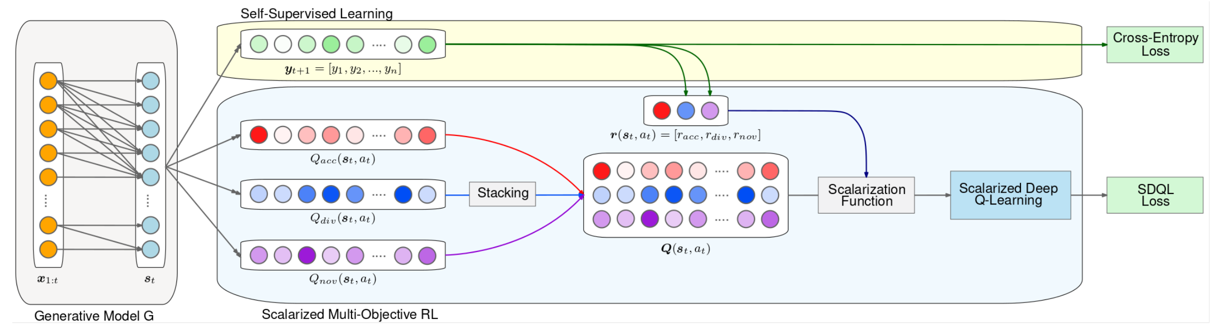


图 1: 推荐系统的 SMORL（SMORL4RS）训练例程，适用于序列或基于会话的推荐系统。生成模型 G 将用户-项目交互序列 $\mathbf{x}_{1:t}$ 映射到潜在状态 $\mathbf{s}_t$。通过使用全连接层，$\mathbf{s}_t$ 被映射到 logits $\mathbf{y}_{t+1}$ 以及一维 Q-值：Q_{acc}、Q_{div} 和 Q_{nov}。多样性和新颖性奖励是通过 logits 获得的最佳预测计算的。向量值 Q-值函数设定为：$\mathbf{Q} = [Q_{\text{acc}}, Q_{\text{div}}, Q_{\text{nov}}]$。SDQL 损失通过标量化函数和 SDQL 算法获得，并与交叉熵损失一起用于训练基模型。

关领域的推荐系统（RS）的方法。我们注意到，通过强化推荐项目的相关性，可以显著阻碍用户探索平台的能力，因为推荐与用户最近的兴趣相似。我们在第 5 节中探讨了这一观点。因此，对于模型来说，学习如何推荐多样化的项目集合以及那些更可能从未被用户发现的项目是至关重要的。

4.2 Reinforcing Diversity

为了使SMORL头部促进多样化的推荐集，我们首先训练一个GRU4Rec模型 [14]，并保存嵌入层 $\mathbf{E}_{\text{div}}$。然后，我们冻结$\mathbf{E}_{\text{div}}$的权重以停止参数的进一步更新。我们将奖励（reward）r_{div}定义为

$$r_{\text{div}} = r_{\text{div}}(\mathbf{s}_t, p_t) = 1 - \cos(l_t, p_t) = 1 - \frac{\mathbf{e}_{l_t}^\top \mathbf{e}_{p_t}}{\|\mathbf{e}_{l_t}\| \|\mathbf{e}_{p_t}\|} \quad (4)$$

在这里，l_t 是会话中最后点击的项目，p_t 是从自监督层获得的最佳预测，而 $\mathbf{e}_x$ 是项目 x 的嵌入，来自于 $\mathbf{E}_{\text{div}}$。我们不使用目前正在训练的模型的嵌入来计算 r_{div}。在训练过程的初始阶段，这会不稳定，从而产生不可靠的多样性奖励。该奖励加强了推荐的多样性，而不仅仅是单个列表。我们考虑基于最佳预测 p_t 和前 k 个推荐的多样性奖励系统，而不仅是最后点击的项目 l_t，但我们并未观察到性能的改善。

4.3 Reinforcing Novelty

给定一个项目，用户可能之前在另一组推荐中见过它，但选择不点击，或者在某个平台上已经遇到过它。因此，在实际应用中，无法跟踪用户可能已经看到的项目，以及推荐那些肯定是新颖的项目。为了解决这个问题并将新颖性和偶然性引入推荐集，我们采取了一种概率方法。不太流行的项目更有可能是新颖的，并导致项目受欢迎程度的更平衡分布。我们使用二元化的项目频率作为我们MORL（多目标强化学习）头的奖励（reward），我们定义如下：

$$r_{\text{nov}} = r_{\text{nov}}(\mathbf{s}_t, p_t) = \begin{cases} 0.0 & p_t \text{ in top } x\% \text{ of most popular items} \\ 1.0 & \text{otherwise} \end{cases}$$

其中 p_t 是从自监督层获得的顶部预测项（top predicted item）。x 的选择基于从训练集推断出的项目流行度的经验分布，即我们将其设置为长尾开始的近似百分位数。此工作中使用的两个数据集具有相似的分布，因此我们设置 $x := 10$。可以看出，准确性奖励（accuracy reward）依赖于会话中的下一个项目，而多样性（diversity）和新颖性（novelty）奖励则依赖于自监督层的顶部预测。

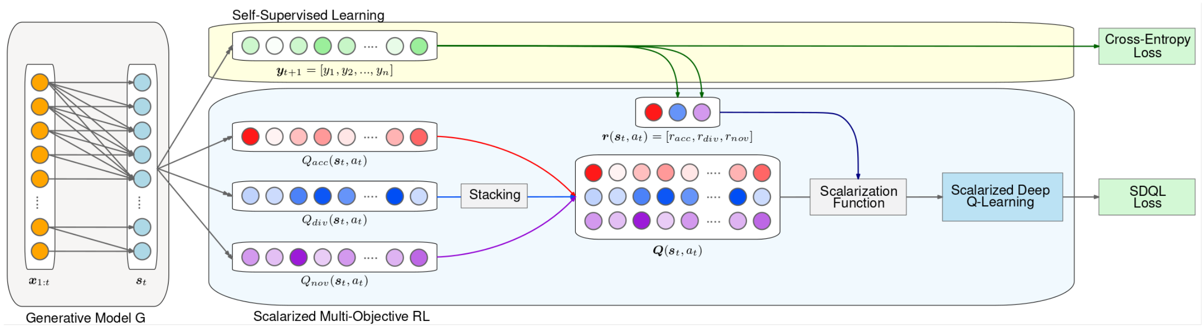


图 1: The SMORL for Recommender Systems (SMORL4RS) training routine for sequential or session-based RS. Generative model G maps user-item interaction sequence $x_{1:t}$ to the latent state s_t. With the use of fully-connected layers, s_t is mapped to logits y_{t+1} and to 1-dimensional Q-values: Q_{acc}, Q_{div}, and Q_{nov}. Diversity and novelty rewards are calculated using the top prediction obtained by the logits. The vector valued Q-value function is set to: $Q = [Q_{acc}, Q_{div}, Q_{nov}]$. SDQL loss is obtained by the scalarization function and SDQL algorithm, and used for training the base model along with the cross-entropy loss.

model G that fine-tunes it by assessing the quality of recommended top item, according to the predefined reward setting and scalarization function f, i.e., importance of objectives.

4.1 Reinforcing Accuracy

For the base model G to learn to provide more relevant recommendations, we expand on [42] and define accuracy reward as

$$r_{acc}(s_t, a_t) = r_{acc}(a_t) = 1, \quad a_t \text{ is a clicked item} \tag{3}$$

From the definition of the reward, the model is rewarded when it matches the next clicked item in the sequence. Xin et. al. [42] suggested using the reward for both clicks and purchases. However, in this work, we introduce a method that can be easily extrapolated from e-commerce to other relevant areas of RS. We note that, by reinforcing the relevance of recommended items, one can significantly hinder the user's ability to explore the platform due to the similarity of the recommendations to the user's recent interests. We explore this claim in Section 5. Therefore, it is crucial for a model to also learn how to recommend diverse sets of items, as well as items that are more probable to never be discovered by the user.

4.2 Reinforcing Diversity

For the SMORL head to promote diverse sets of recommendations, we first train a GRU4Rec model [14], and save the embedding layer $\mathbf{E}_{div}$. We then freeze the weights of $\mathbf{E}_{div}$ to stop further updates of the parameters. We define the reward r_{div} as

$$r_{div} = r_{div}(s_t, p_t) = 1 - \cos(l_t, p_t) = 1 - \frac{\mathbf{e}_{l_t}^\top \mathbf{e}_{p_t}}{\|\mathbf{e}_{l_t}\| \|\mathbf{e}_{p_t}\|} \tag{4}$$

where l_t is the last clicked item in the session, p_t is a top prediction obtained from self-supervised layer, and $\mathbf{e}_x$ is the embedding of the item x, obtained from $\mathbf{E}_{div}$. We do not use the embedding of a model that is currently trained for calculation r_{div}. It would be unstable at the beginning of the training process, which would produce unreliable diversity rewards. This reward reinforces diversity across a session of recommendations rather than just over a single slate. Basing the diversity reward system on top prediction p_t and top-k recommendations instead of only

Algorithm 1: Training Procedure of SMORL

Input : user-item interaction sequence set $\mathcal{X}$, recommendation model G, SMORL head $\mathbf{Q}$, supervised head S, predefined parameters α and $\mathbf{w}$

Output : all parameters in the learning space Θ

```
Initialize all trainable parameters
Create G′, Q′, as copies of G and Q, respectively
repeat
    Draw a mini-batch of (x_{1:t}, a_t) from X
    s_t = G(x_{1:t}), s′_t = G′(x_{1:t})
    s_{t+1} = G(x_{2:t+1}), s′_{t+1} = G′(x_{2:t+1})
    Generate random variable z ∈ (0, 1)
     uniformly
    if z < 0.5 then
        a* = argmax_a [Q(s_{t+1}, a) · w]
        pred = argmax S(s_t)
        Set reward r_t = stack(r_acc, r_div, r_nov)
        L_SDQL = w^T (r_t + γQ′(s′_{t+1}, a*) − Q(s_t, a_t))^2
        Calculate L_s
        L_SMORL = L_s + α L_SDQL
        Perform updates by ∇_Θ L_SMORL
    else
        a* = argmax_a [Q′(s′_{t+1}, a) · w]
        pred = argmax S(s_t)
        Set reward r_t = stack(r_acc, r_div, r_nov)
        L_SDQL =
         (w^T (r_t + γQ(s_{t+1}, a*) − Q′(s′_t, a_t))^2
        Calculate L_s
        L_SMORL = L_s + α L_SDQL
        Perform updates by ∇_Θ L_SMORL
    end
until converge;
Return all parameters in Θ
```

4.4 Scalarized Multi-Objective RL for RS

推荐本质上是一个多目标问题，因此，库存自我监督学习（self-supervised learning），甚至单目标强化学习（RL）方法都不能满足所有理想（或必要）目标。我们将提出的三个目标整合到一个单一的SMORL方法中，在每个时间戳下找到一个最佳动作（动作），该动作根据预定义的用户效用函数，或在此情况下，依据式(2)中w的配置来考虑所有目标。SMORL高度可定制并适应特定提供者的目标——可以定义不同的奖励（奖励）系统，从而导致一个提供更相关、新颖、多样、意外或偶然推荐的RS。我们优化的最终损失是：

$$L_{SMORL} = L_s + \alpha L_{SDQL} \tag{5}$$

在这里，L_s 是交叉熵损失，而 α 是一个超参数，使我们能够控制 SMORL 部分的影响。为了增强学习的稳定性，我们交替训练两个可学习参数的副本。算法 1 描述了 SMORL 的训练过程。值得注意的是，在训练完成后，仅使用基础模型的自监督部分来生成推荐，同时通过 SMORL 部分观察与不同指标相关的效果。

这一训练框架可以集成到现有的推荐模型中，只要它们遵循之前讨论的通用架构。这适用于近年来引入的大多数基于会话或顺序的推荐模型。在这项工作中，我们使用交叉熵损失用于自监督部分，但其他模型可以结合不同的损失函数 [13, 30]。

此外，SMORL 是一个高度模块化的框架，用户可以重新加权和“停用”特定的强化学习 (RL) 目标，或在精心设计的奖励 (reward) 方案的帮助下添加更多的目标。最终，这一机制使推荐系统 (RS) 能够专注于提供者的特定短期和长期目标。然而，我们的实验结果表明，在大多数情况下，由三种 RL 目标正则化的模型在所有质量指标上表现最佳。

the last clicked item l_t was considered, but we observed no improvement in performance.

4.3 Reinforcing Novelty

Given an item, a user may have previously seen it in another set of recommendations but chose not to click it, or already encountered it on some other platform. Therefore, in a real-world use case, it is impossible to track the items that a user may have already seen, and to suggest items that are certain to be novel. To address this issue and introduce novelty and serendipity into the set of recommendations, we take a probabilistic approach. Less popular items are more likely to be novel and lead to a more balanced distribution of item popularity. We use binarized item frequency as a novelty reward for our MORL head, which we define as follows:

$$r_{\text{nov}} = r_{\text{nov}}(\mathbf{s}_t, p_t) = \begin{cases} 0.0 & p_t \text{ in top } x\% \text{ of most popular items} \\ 1.0 & \text{otherwise} \end{cases}$$

where p_t is the top predicted item obtained from the self-supervised layer. The choice of x is based on the empirical distribution of the item popularity inferred from the training set, i.e., we set it to the approximate percentile where the long tail starts. Both datasets used in this work have a similar distribution, so we set $x := 10$. As can be seen, accuracy reward depends on the next item in the session, while diversity and novelty rewards depends on the top prediction from self-supervised layer.

4.4 Scalarized Multi-Objective RL for RS

Recommendation is by nature a multi-objective problem and, as such, stock self-supervised learning, or even single-objective RL methods, cannot satisfy all desirable (or necessary) goals. We integrate the three proposed objectives into a single SMORL method that at each timestamp finds

Algorithm 1: Training Procedure of SMORL

Input : user-item interaction sequence set $\mathcal{X}$, recommendation model G, SMORL head $\mathbf{Q}$, supervised head S, predefined parameters α and $\mathbf{w}$

Output: all parameters in the learning space Θ

Initialize all trainable parameters
Create G', $\mathbf{Q}'$, as copies of G and $\mathbf{Q}$, respectively
repeat
  Draw a mini-batch of $(\mathbf{x}_{1:t}, a_t)$ from $\mathcal{X}$
  $\mathbf{s}_t = G(\mathbf{x}_{1:t}), \mathbf{s}'_t = G'(\mathbf{x}_{1:t})$
  $\mathbf{s}_{t+1} = G(\mathbf{x}_{2:t+1}), \mathbf{s}'_{t+1} = G'(\mathbf{x}_{2:t+1})$
  Generate random variable $z \in (0, 1)$ uniformly
  if $z < 0.5$ **then**
    $a^* = \text{argmax}_a [\mathbf{Q}(\mathbf{s}_{t+1}, a) \cdot \mathbf{w}]$
    pred = argmax $\text{S}(\mathbf{s}_t)$
    Set reward $\mathbf{r}_t = \text{stack}(r_{\text{acc}}, r_{\text{div}}, r_{\text{nov}})$
    $L_{\text{SDQL}} = \mathbf{w}^\top(\mathbf{r}_t + \gamma\mathbf{Q}'(\mathbf{s}'_{t+1}, a^*) - \mathbf{Q}(\mathbf{s}_t, a_t))^2$
    Calculate L_s
    $L_{\text{SMORL}} = L_s + \alpha L_{\text{SDQL}}$
    Perform updates by $\nabla_\Theta L_{\text{SMORL}}$
  else
    $a^* = \text{argmax}_a [\mathbf{Q}'(\mathbf{s}'_{t+1}, a) \cdot \mathbf{w}]$
    pred = argmax $\text{S}(\mathbf{s}_t)$
    Set reward $\mathbf{r}_t = \text{stack}(r_{\text{acc}}, r_{\text{div}}, r_{\text{nov}})$
    $L_{\text{SDQL}} = (\mathbf{w}^\top(\mathbf{r}_t + \gamma\mathbf{Q}(\mathbf{s}_{t+1}, a^*) - \mathbf{Q}'(\mathbf{s}'_t, a_t))^2$
    Calculate L_s
    $L_{\text{SMORL}} = L_s + \alpha L_{\text{SDQL}}$
    Perform updates by $\nabla_\Theta L_{\text{SMORL}}$
  end
until *converge*;
Return all parameters in Θ

an optimal action that takes into consideration all objectives according to a predefined user utility function, or in

5 EXPERIMENTS

我们报告了我们实验的结果[3] 在两个真实世界的连续电子商务数据集上。对于所有基础模型，我们使用自监督头生成推荐。我们解决以下研究问题：

RQ1: 当集成时，所提出的方法是否提高了基础模型的性能？

RQ2: 我们能否控制准确性、多样性和新颖性之间的平衡？

RQ3: 我们能否通过调整SMORL部分的梯度强度来增加其影响力？

5.1 Experimental Settings

5.1.1 Datasets: RC15[4] 和 RetailRocket[5]，表 1。

RC15. 此数据集基于 2015 年 RecSys 挑战（RecSys Challenge 2015）。该数据集是基于会话的，每个会话包含一系列的点击和购买[6]。我们丢弃长度小于 3 的会话，然后抽样 200K 个会话的子集。

RetailRocket. 此数据集是从一个真实的电子商务网站收集的。它包含查看和加入购物车的会话事件。为了与 RC15 数据集保持一致，我们将查看视为点击。我们删除互动次数少于三次（3）的项目，以及长度小于三（3）的序列。

5.1.2 Quality of Recommendation Metrics.

准确性度量。 推荐项目集的相关性通常通过两个度量来评估：命中率（HR）和归一化折扣累积增益（NDCG）。HR@k是一种基于召回的度量，衡量真实项是否出现在推荐列表的前k个位置。我们将点击的HR定义为：

$$\text{HR(click)} = \frac{\#\text{hits among clicks}}{\#\text{clicks}}$$

另一方面，NDCG是一种对排名敏感的度量，它对推荐列表中的顶级位置赋予更高的分数[18]

[3]实现可以在 https://drive.google.com/file/d/1lVeKlajOkZ4n9Rl2VmJvYR9i1aXWkR2j/view?usp=sharing 找到

[4]https://recsys.acm.org/recsys15/challenge/

[5]https://www.kaggle.com/retailrocket/ecommerce-dataset

[6]在本研究中,我们仅考虑点击。

表 1: Dataset statistics.

Dataset	RC15	RetailRocket
#sequences	200,000	195,523
#items	26,702	70,852
#clicks	1,110,965	1,176,680
#purchase	43,946	57,269

多样性与新颖性指标。 推荐系统中的多样性可以在个体或整体层面上进行查看。例如，如果推荐系统向所有用户提供相同的一组十个不同的项目，则每个用户的推荐列表会是多样的，即它具有较高的个体多样性。然而，系统只能从整个项目池中推荐十个项目，因此整体多样性可能微不足道。因此，在我们的实验中，我们使用Coverage@k (CV@k)来测量整体多样性，$k \in \{1, 5, 10, 20\}$。更具体地说，我们在两个集合上测量CV@k：所有项目的集合和不太受欢迎项目的集合。覆盖率可以计算为所有顶级-k推荐的验证或测试序列所覆盖的所有项目（不太受欢迎项目）的百分比。

推荐的重复性。 我们引入重复性（Repetitiveness, R），这是一种评估推荐有用性的新的度量。我们认为该指标是判断推荐系统如何易于创建过滤泡沫的良好替代，因为它测量了推荐列表中顶级-k位置的每个会话平均重复次数。我们测量R@k，$k \in \{5, 10, 20\}$，并将其定义为：

$$R@K = \frac{1}{N}\sum_{i=1}^{N} \#\text{repetitions in top-}k\text{ items of session } i \tag{6}$$

其中 N 是测试（或验证）集中的会话总数。

5.1.3 Evaluation Protocols. 我们使用 5 折交叉验证 (5-fold cross-validation) 进行性能评估，训练、验证和测试的比例为 8:1:1。我们报告所有折叠的平均性能。

5.1.4 Baselines. 我们在四个最先进的 (state-of-the-art) 生成 (generative) 序列推荐模型中集成了 SMORL：

- GRU4Rec [14]: 该方法使用GRU（门控递归单元）对输入序列进行建模。GRU4Rec的最终隐

this case, according to the configuration of $\mathbf{w}$ from Eq.(2). SMORL is highly customizable and adaptable to a specific provider's goals - one can define different reward systems that can result in a RS that provides more relevant, novel, diverse, unexpected, or serendipitous recommendations. The final loss that we optimize is:

$$L_{SMORL} = L_s + \alpha L_{\mathrm{SDQL}} \tag{5}$$

where L_s is a cross-entropy loss, and α is a hyperparameter that enables us to control the influence of SMORL part. In order to enhance the learning stability, we alternately train two copies of learnable parameters. Algorithm 1 describes the training procedure of SMORL. It should be noted that after the training is finished, only the self-supervised part of the base model is used to produce recommendations, while the effects with respect to different metrics are observed from the regularization by the SMORL part.

This training framework can be integrated in existing recommendation models, provided they follow the general architecture discussed earlier. This is the case for most session-based or sequential recommendation models introduced over the last years. In this work, we use the cross-entropy loss for the self-supervised part but other models can incorporate different loss functions [13, 30].

In addition, SMORL is a highly modular framework, where one can re-weight and "deactivate" specific RL objectives, or add more of them with the help of a carefully designed reward schema. Ultimately, this mechanism allows the RS to focus on providers' specific short-term and long-term goals. However, our experimental results show that models regularized by all three RL objectives perform the best in most cases, with respect to all quality metrics.

5 EXPERIMENTS

We report the results of our experiments[3] on two real-world sequential e-commerce datasets. For all base models, we used the self-supervised head to generate recommendations. We address the following research questions:

RQ1: When integrated, does the proposed method increase the performance of the base models?

RQ2: Can we control the balance between accuracy, diversity and novelty?

RQ3: Can we increase the influence of SMORL part by adjusting the intensity of its gradient?

5.1 Experimental Settings

5.1.1 Datasets: RC15[4] and RetailRocket[5], Table 1.

RC15. This dataset is based on the RecSys Challange 2015. The dataset is session-based and each session contains a sequence of clicks and purchases[6]. We discard sessions whose length is smaller than 3 and then sample a subset of 200K sessions.

RetailRocket. This dataset is collected from a real-world e-commerce website. It contains session events of viewing and adding to cart. To keep in line with the RC15 dataset, we treat views as clicks. We remove the items which are interacted less than three times (3), and the sequences whose length is smaller than three (3).

5.1.2 Quality of Recommendation Metrics.

Accuracy metrics. Relevance of the recommended item set is usually measured with two metrics: hit ration (HR) and normalized discounted cumulative gain (NDCG). HR@k is a recall-based metric, measuring whether the

[3]The implementation can be found at https://drive.google.com/file/d/1lVeKlajOkZ4n9Rl2VmJvYR9i1aXWkR2j/view?usp=sharing
[4]https://recsys.acm.org/recsys15/challenge/
[5]https://www.kaggle.com/retailrocket/ecommerce-dataset
[6]In this work, we only consider clicks.

藏状态被视为输入序列的潜在表征（latent representation）。

- Caser [36]：这种最近引入的基于卷积神经网络（CNN，Convolutional Neural Network）的方法通过对前一个项目的嵌入矩阵（embedding matrix）应用卷积操作来捕捉序列信号。
- NextItNet [43]：该方法通过使用膨胀卷积神经网络（dilated CNN）来扩大感受野，并使用残差连接（residual connection）来增加网络深度，从而增强了Caser。
- SASRec [19]：该基准模型受到自注意力（self-attention）的启发，并使用transformer [37]架构对用户-物品交互序列进行编码。transformer编码器的输出被视为潜在表征（latent representation）。

5.1.5 Parameter settings. 对于这两个数据集，输入序列由目标时间戳之前的最后 10 个项目组成。如果序列长度小于 10，我们用一个填充项进行补充。我们使用 Adam 优化器（optimizer）训练所有模型 [20]。小批量（mini-batch）大小设定为 256。学习率对于 RC15 设置为 0.01，对于 RetailRocket 设置为 0.005。我们在 RC15 的每 5, 000 批更新和 RetailRocket 的每 10, 000 批更新上评估验证集。为了确保公平比较，所有模型的项目嵌入（embedding）大小均设置为 64。对于 GRU4Rec 模型，隐状态（hidden state）的大小设置为 64。对于 Caser，我们使用一个垂直卷积（vertical convolution）滤波器和 16 个水平（horizontal）滤波器，其高度设置为 $\{2, 3, 4\}$。丢弃率（drop-out ratio）设置为 0.1。对于 NextItNet，我们使用作者报告的相同参数。对于 SASRec，自注意力（self-attention）中的头数设置为 1，根据其原始论文 [19]。我们将折扣因子（discount factor）γ 设置为 0.5，如 Xin et al. [42] 推荐的那样。

5.2 Performance Comparison (RQ1)

对于这两个数据集，SQN 方法 [42] 在向用户推荐相关项目方面优于基线模型。然而，通过提高基线模型的准确性，它导致模型在多样性（diversity）和新颖性（novelty）上出现"漂移"。这导致基线模型的覆盖率指标显著下降（最高可达 20%），无论是对所有项目还是对不太受欢迎的项目。结合这一事实，推荐的重复性增加表明，仅仅强化准确性可能明显影响感知的体验质量。此外，显然应该同时优化模型以实现多样性和新颖性，从而在相对指标之间取得平衡。在表 2 和表 3 中，我们看到通过使用 SMORL 方法，我们不仅在准确性、多样性和新颖性之间取得平衡，而且在所有指标上持续优于相应的基线模型，并在某种程度上提高了它们的准确性。与基线模型相比，多样性和新颖性的提高高达 20%，而与 SQN 模型相比，提升高达 40%。基线模型准确性的提高可以归因于大多数用户的兴趣多样，推荐系统（RS）产生的推荐无法满足这些兴趣 [1]。图 2 显示了在 RC15 测试集上获得的累计多样性和新颖性奖励的差异。当使用 SMORL 框架训练基础模型时，我们注意到累计多样性和新颖性奖励显著增加。此外，表 2 和表 3 中的结果表明，强化多样性和新颖性在这些指标上引入了显著的改善，这些指标与感知的体验质量和参与度高度相关。

5.3 Reinforcing a Subset of Objectives (RQ2)

使用SMORL的一个优点是其目标平衡能力，它通过在公式(2)中使用不同配置的$\mathbf{w}$对目标进行重新加权。在我们的设置中，$\mathbf{w}$的第一个条目对应于准确性目标的权重，第二个对应于多样性，第三个对应于新颖性目标。我们使用以下参数$\mathbf{w}$的配置进行实验：

$$\mathbf{w} \in \{(0, 1, 0), (0, 0, 1), (0, 1, 1), (1, 1, 0), (1, 0, 1)\} \tag{7}$$

在这里，我们旨在展示在强化三项重要目标的子集时性能的差异。我们在此分析中不包括 $\mathbf{w} = (1, 0, 0)$， ground-truth item is in the top-k positions of the recommendation list. We define HR for clicks as:

$$\text{HR(click)} = \frac{\text{\#hits among clicks}}{\text{\#clicks}}$$

On the other hand, NDCG is a rank sensitive metric that assigns higher scores to top positions in the recommendation list [18]

Diversity & Novelty metrics. Diversity in RS can be viewed at either individual or aggregate level. For example, if the RS was to provide the same set of ten dissimilar items to all users, the recommendation list for each user would be diverse, i.e., it would have high individual diversity. However, the system can only recommend ten items out of the entire item pool and, thus, the aggregate diversity would be negligible. Therefore, in our experiments, we measure aggregate diversity using Coverage@k (CV@k), $k \in \{1, 5, 10, 20\}$. More specifically, we measure CV@k on two sets: set of all items and a set of less popular items. Coverage can be computed as percentage of all items (less popular items) covered by all top-k recommendations of the validation or test sequences.

Repetitiveness of Recommendations. We introduce Repetitiveness (R), a novel metric for evaluating the usefulness of recommendations. We consider this metric a good proxy as to how easily a RS can create a filter bubble, as it measures the per session average of repetitions in the top-k positions of recommendations lists. We measure R@k, $k \in \{5, 10, 20\}$ and define it as:

$$R@K = \frac{1}{N} \sum_{i=1}^{N} \text{\#repetitions in top-}k\text{ items of session } i \tag{6}$$

where N is the total number of sessions in test (or validation) set.

5.1.3 Evaluation Protocols. We use 5-fold cross-validation for our performance evaluation, with a ratio of 8:1:1 for training, validation, and testing. We report average performance across all folds.

表 1: Dataset statistics.

Dataset	RC15	RetailRocket
#sequences	200,000	195,523
#items	26,702	70,852
#clicks	1,110,965	1,176,680
#purchase	43,946	57,269

5.1.4 Baselines. We integrated SMORL in four state-of-the-art (generative) sequential recommendation models:

- GRU4Rec [14]: This method uses a GRU to model the input sequences. The final hidden state of the GRU4Rec is treated as the latent representation for the input sequence.
- Caser [36]: This recently introduced CNN-based method captures sequential signals by applying convolution operations on the embedding matrix of previous items.
- NextItNet [43]: This method enhances Caser by using dilated CNN to enlarge the receptive field and residual connection to increase the network depth.
- SASRec [19]: This baseline is motivated from self-attention and uses the Transformer [37] architecture to encode sequences of user-item interactions. The output of the Transformer encoder is treated as the latent representation.

5.1.5 Parameter settings. For both datasets the input sequences comprise of the last 10 items before the target timestamp. If the sequence length is less than 10, we complement it with a padding item. We train all models with the Adam optimizer [20]. The mini-batch size is set as 256. The learning rate is set as 0.01 for RC15 and 0.005 for RetailRocket. We evaluate on the validation set every 5,000 batches of updates on RC15, and every 10,000 batches of updates on RetailRocket. To ensure a fair comparison, the item embedding size is set as 64 for all models. For the GRU4Rec model, the size of the hidden state is set as

Woodstock '18, June 03–05, 2018, Woodstock, NY Dušan Stamenković, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas

表 2: 在RC15数据集上的推荐性能。NG是NDCG。CV是覆盖率。粗体字表示最高得分。

Models	accuracy				diversity				novelty				repetitivenes		
	HR@10	NG@10	HR@20	NG@20	CV@1	CV@5	CV@10	CV@20	CV@1	CV@5	CV@10	CV@20	R@5	R@10	R@
GRU	0.3793	0.2279	0.4581	0.2478	0.2481	0.4330	0.5188	0.5942	0.1777	0.3707	0.4654	0.5492	12.11	25.63	53
GRU-SQN	0.3946	0.2394	0.4741	0.2587	0.2406	0.4025	0.4710	0.5364	0.1656	0.3363	0.4122	0.4849	12.20	25.81	53
GRU-SMORL	**0.4007**	**0.2433**	**0.4793**	**0.2632**	**0.2825**	**0.4758**	**0.5577**	**0.6334**	**0.2086**	**0.4176**	**0.5086**	**0.5927**	**11.29**	**23.81**	**48**
Caser	0.3593	0.2177	0.4371	0.2372	0.2631	0.4349	0.5019	0.5608	0.1912	0.3724	0.4466	0.5120	14.38	29.65	60
Caser-SQN	**0.3668**	0.2223	**0.4448**	**0.2420**	0.2154	0.3525	0.4057	0.4557	0.1411	0.2810	0.2154	0.3953	14.45	29.79	60
Caser-SMORL	0.3664	**0.2224**	0.4425	0.2417	**0.3174**	**0.5157**	**0.5944**	**0.6685**	**0.2476**	**0.4621**	**0.5495**	**0.6316**	**13.77**	**28.56**	**58**
NtItNet	0.3885	0.2332	0.4684	0.2535	0.2950	0.4914	0.5705	0.6427	0.2313	0.4354	0.5228	0.6030	10.03	22.02	46
NtItNet-SQN	0.4083	0.2492	0.4878	0.2693	0.2737	0.4572	0.5183	0.5715	0.2082	0.3975	0.4649	0.5239	10.19	22.32	47
NtItNet-SMORL	**0.4116**	**0.2505**	**0.4898**	**0.2703**	**0.3385**	**0.5639**	**0.6518**	**0.7283**	**0.2720**	**0.5156**	**0.6131**	**0.6981**	**9.97**	**21.73**	**45**
SASRec	0.4257	0.2599	0.5053	0.2801	0.2971	0.5208	0.6046	0.6792	0.2298	0.4679	0.5607	0.6436	10.62	23.24	49
SASRec-SQN	0.4288	0.2630	0.5073	0.2829	0.2701	0.4527	0.5194	0.5755	0.2018	0.3922	0.4660	0.5283	10.94	23.85	50
SASRec-SMORL	**0.4315**	**0.2651**	**0.5104**	**0.2851**	**0.3380**	**0.5755**	**0.6508**	**0.7158**	**0.2698**	**0.5285**	**0.6120**	**0.6842**	**10.38**	**22.79**	**48**

表 3: 在RetailRocket数据集上的推荐性能。NG是NDCG。CV是覆盖率（Coverage）。粗体字表示最高分。

Models	accuracy				diversity				novelty				repetitivenes		
	HR@10	NG@10	HR@20	NG@20	CV@1	CV@5	CV@10	CV@20	CV@1	CV@5	CV@10	CV@20	R@5	R@10	R
GRU	0.2673	0.1878	0.3082	0.1981	0.2439	0.4695	0.5699	0.6632	0.1837	0.4139	0.5238	0.6267	14.25	29.44	6
GRU-SQN	0.2967	0.2094	0.3406	0.2205	0.2180	0.4114	0.4975	0.5763	0.1526	0.3489	0.4430	0.5299	14.62	30.19	6
GRU-SMORL	**0.3060**	**0.2103**	**0.3535**	**0.2224**	**0.2796**	**0.5369**	**0.6419**	**0.7353**	**0.2154**	**0.4871**	**0.6029**	**0.7064**	**13.53**	**28.02**	**5**
Caser	0.2302	0.1675	0.2628	0.1758	0.2327	0.4379	0.5133	0.5718	0.1643	0.3773	0.4605	0.5252	16.16	33.24	6
Caser-SQN	0.2454	0.1778	0.2803	0.1867	0.2088	0.3880	0.4511	0.5021	0.1387	0.3219	0.3914	0.4479	16.88	34.50	7
Caser-SMORL	**0.2657**	**0.1898**	**0.3052**	**0.1998**	**0.2855**	**0.5411**	**0.6324**	**0.7138**	**0.2224**	**0.4917**	**0.5925**	**0.6827**	**15.90**	**32.47**	**6**
NtItNet	0.3007	0.2060	0.3506	0.2186	0.2867	0.5113	0.6033	0.6837	0.2305	0.4595	0.5605	0.6495	12.25	25.76	5
NtItNet-SQN	0.3129	0.2150	0.3586	0.2266	0.2802	0.5255	0.6077	0.6750	0.2184	0.4747	0.5651	0.6395	12.27	25.93	5
NtItNet-SMORL	**0.3183**	**0.2222**	**0.3659**	**0.2342**	**0.3429**	**0.6335**	**0.7351**	**0.8129**	**0.2800**	**0.5938**	**0.7062**	**0.7924**	**10.92**	**22.89**	**4**
SASRec	0.3085	0.2107	0.3572	0.2227	0.2767	0.5305	0.6300	0.7149	0.2171	0.4806	0.5899	0.6838	15.67	32.27	6
SASRec-SQN	0.3302	0.2279	0.3803	0.2406	0.2393	0.4617	0.5490	0.6254	0.1753	0.4040	0.5001	0.5847	15.60	32.20	6
SASRec-SMORL	**0.3521**	**0.2477**	**0.4028**	**0.2605**	**0.3037**	**0.5724**	**0.6672**	**0.7476**	**0.2366**	**0.5261**	**0.6311**	**0.7202**	**12.58**	**26.69**	**5**

因为 SMORL 与 [42] 中的 SQN 方法等价，我们的结果在所有模型中表现出完全相同的行为。

我们在本工作中涉及的目标之间存在复杂的关系。例如，训练过程开始时的相关性和多样性是相关的，即，更多样化的推荐会产生更相关的推荐，而随着训练的进展，它们之间的相关性变为负数。多样性和新颖性是交织在一起的目标，例如，一组多样化的推荐项目更可能包含新颖的项目。另一方面，项目的受欢

64. For Caser, we use one vertical convolution filter and 16 horizontal filters, whose heights are set from $\{2, 3, 4\}$. The drop-out ratio is set as 0.1. For NextItNet, we use the same parameters reported by authors. For SASRec, the number of heads in self-attention is set as 1, according to its original paper [19]. We set the discount factor γ to 0.5, as recommended by Xin et al. [42].

5.2 Performance Comparison (RQ1)

For both datasets, the SQN method [42] outperforms the baselines with respect to recommending relevant items to the users. However, by increasing the accuracy of the baseline model, it causes it to "drift" from diversity and novelty. This results in a substantial decrease (up to 20%) of coverage metrics for the baseline model, both on all and less popular items. Together with this fact, increased repetitiveness of recommendations suggests that reinforcing accuracy alone may hinder significantly the perceived quality of experience. Furthermore, it is evident that one should simultaneously optimize the model towards diversity and novelty to achieve a balance between opposing metrics. In Table 2 and Table 3, we see that by using the SMORL method we not only obtain a balance between accuracy, diversity and novelty, but we consistently outperform the corresponding baselines across all metrics and, to some extent, we also improve their accuracy power. The increase in diversity and novelty is up to 20% relative to the baseline model, and up to 40% relative to the SQN model. Increases in the accuracy of the baseline models can be attributed to most users having diverse interests that cannot be satisfied by the recommendations produced by an RS [1]. Figure 2 displays the difference in cumulative diversity and novelty rewards obtained on the RC15 test set. When a base model is trained with the SMORL framework, we note a significant increase in the cumulative diversity and novelty rewards. Also, the results in Tables 2 and 3 suggest that reinforcing diversity and novelty introduces a notable improvement in these metrics, which are highly correlated with perceived quality of experience and engagement.

5.3 Reinforcing a Subset of Objectives (RQ2)

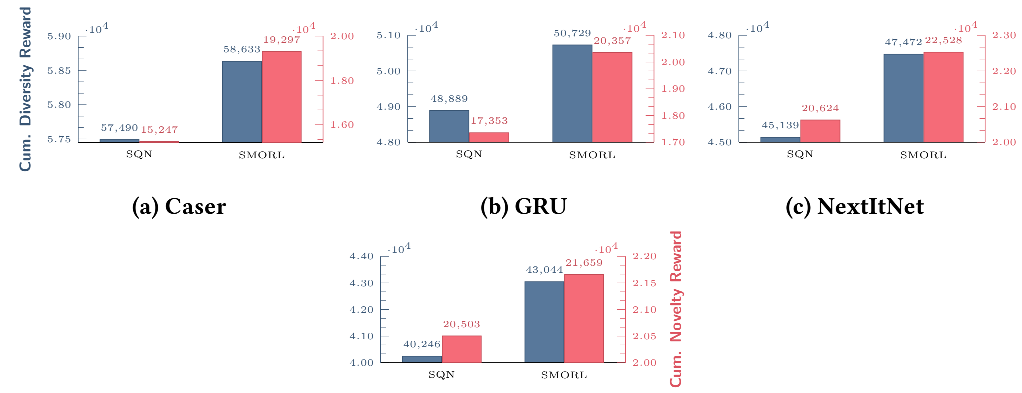
One of the advantages of using SMORL is its objective-balancing capability, which works by re-weighting the objectives using different configurations of $\mathbf{w}$'s in Eq.(2). In our setting, the first entry of $\mathbf{w}$ corresponds to the strength of accuracy objective, the second to diversity, and the third to novelty objective. We conduct experiments with the following configurations of the parameter $\mathbf{w}$:

$$\mathbf{w} \in \{(0,1,0), (0,0,1), (0,1,1), (1,1,0), (1,0,1)\} \quad (7)$$

Here, we aim to demonstrate the difference in performance when reinforcing a subset of three important objectives. We do not include $\mathbf{w} = (1, 0, 0)$ in this analysis, since SMORL becomes equivalent to SQN method from [42] and our results show exactly the same behaviour across all models.

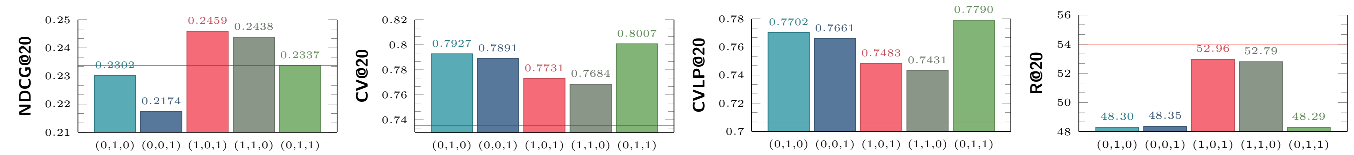
The objectives that we address in this work have a complex relationship. For example, relevance and diversity at the beginning of the training process are correlated, i.e., more diverse recommendations produce more relevant recommendations, while their correlation becomes negative as the training progresses. Diversity and novelty are intertwined objectives, e.g., a diverse set of recommended items is more likely to contain novel items. On the other hand, the popularity of items follows a power distribution and, therefore, less popular items make up to 90% of the dataset, which means that items likely to be novel are inherently diverse. Given that the proposed method is not a pure MORL model, but rather a regularizer that forces the base model to capture different (and often competing) objectives, the intricacies of optimizing and balancing multiple objectives pose a significant research

Choosing the Best of Both Worlds:
Diverse and Novel Recommendations through Multi-Objective Reinforcement Learning Woodstock '18, June 03–05, 2018, Woodstock, NY



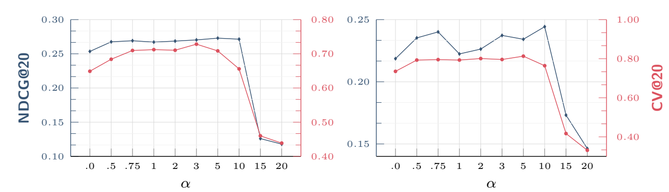
(a) Caser (b) GRU (c) NextItNet (d) SASRec

图 2: 在SMORL和SQN框架下，对基模型在RC15数据集上的累计奖励（cumulative rewards）进行比较。



(a) Accuracy (b) Diversity (c) Novelty (d) Repetitiveness

图 3: 当增强目标子集的性能比较（通过使用来自 Eq.(7）的不同配置的 w 实现）。红线表示股票 NextItNet 模型的相关指标 - 未使用 SMORL4RS 框架进行训练。



(a) (b)

图 4: 在RC15 (4a) 和RetailRocket (4b) 数据集上具有不同SMORL梯度强度的NextItNet

迎程度遵循一个幂律分布，因此，受欢迎程度较低的项目占数据集的 90%，这意味着很可能是新颖的项目本质上是多样的。考虑到所提出的方法不是一个纯粹的 MORL 模型，而是强制基模型捕捉不同（并且通常是相互竞争的）目标的正则化器，优化和平衡多个目标的复杂性构成了一个重大的研究挑战。在本节中，我们的目标是展示我们可以控制每个目标的影响程度，而不是如何找到理想的平衡。通过这种控制的能力，出现了许多工程可能性，例如部署多个 SMORL4RS 智能体，并在线决定用户是否应从一个旨在新颖性、多样性或准确性优化的智能体中接收推荐。

图 3 显示了由 SMORL 智能体 (agent) 使用上述权重配置 w 对 RetailRocket 数据集进行正规化的 NextItNet-SMORL 模型的比较，而对于 RC15 数据集和其他模型也可以观察到类似的行为。更具体地说，图 3a 指出，如果我们仅针对新颖性对模型进行正规化，我们将牺牲其推荐相关项目的能力。如果我们仅强化多样性，这种现象也是存在的，但 NDCG@20 指标的下降并不明显。另一方面，如果我们共同优化多样性和新颖性，我们并未观察到基础模型的准确性下降。此外，如果我们将准确性目标包含在其他两个目标之中，我们观察到相关指标的增加。从图 3b 和 3c 中，我们注意到包含准确性目标会以牺牲多样性和新颖性为代价，而共同优化多样性和新颖性则在这些指标方面产生最佳结果。类似地，通过包含准确性目标，我们与优化多样性和新颖性组合的 NextItNet-SMORL 模型相比，增加了重复性。

表 2: Recommendation performance on RC15 dataset. NG is NDCG. CV is Coverage. Boldface denotes highest score.

Models	accuracy				diversity				novelty				repetitivenes		
	HR@10	NG@10	HR@20	NG@20	CV@1	CV@5	CV@10	CV@20	CV@1	CV@5	CV@10	CV@20	R@5	R@10	R@
GRU	0.3793	0.2279	0.4581	0.2478	0.2481	0.4330	0.5188	0.5942	0.1777	0.3707	0.4654	0.5492	12.11	25.63	53
GRU-SQN	0.3946	0.2394	0.4741	0.2587	0.2406	0.4025	0.4710	0.5364	0.1656	0.3363	0.4122	0.4849	12.20	25.81	53
GRU-SMORL	**0.4007**	**0.2433**	**0.4793**	**0.2632**	**0.2825**	**0.4758**	**0.5577**	**0.6334**	**0.2086**	**0.4176**	**0.5086**	**0.5927**	**11.29**	**23.81**	**48**
Caser	0.3593	0.2177	0.4371	0.2372	0.2631	0.4349	0.5019	0.5608	0.1912	0.3724	0.4466	0.5120	14.38	29.65	60
Caser-SQN	**0.3668**	0.2223	**0.4448**	**0.2420**	0.2154	0.3525	0.4057	0.4557	0.1411	0.2810	0.2154	0.3953	14.45	29.79	60
Caser-SMORL	0.3664	**0.2224**	0.4425	0.2417	**0.3174**	**0.5157**	**0.5944**	**0.6685**	**0.2476**	**0.4621**	**0.5495**	**0.6316**	**13.77**	**28.56**	**58**
NtItNet	0.3885	0.2332	0.4684	0.2535	0.2950	0.4914	0.5705	0.6427	0.2313	0.4354	0.5228	0.6030	10.03	22.02	46
NtItNet-SQN	0.4083	0.2492	0.4878	0.2693	0.2737	0.4572	0.5183	0.5715	0.2082	0.3975	0.4649	0.5239	10.19	22.32	47
NtItNet-SMORL	**0.4116**	**0.2505**	**0.4898**	**0.2703**	**0.3385**	**0.5639**	**0.6518**	**0.7283**	**0.2720**	**0.5156**	**0.6131**	**0.6981**	**9.97**	**21.73**	**45**
SASRec	0.4257	0.2599	0.5053	0.2801	0.2971	0.5208	0.6046	0.6792	0.2298	0.4679	0.5607	0.6436	10.62	23.24	49
SASRec-SQN	0.4288	0.2630	0.5073	0.2829	0.2701	0.4527	0.5194	0.5755	0.2018	0.3922	0.4660	0.5283	10.94	23.85	50
SASRec-SMORL	**0.4315**	**0.2651**	**0.5104**	**0.2851**	**0.3380**	**0.5755**	**0.6508**	**0.7158**	**0.2698**	**0.5285**	**0.6120**	**0.6842**	**10.38**	**22.79**	**48**

表 3: Recommendation performance on RetailRocket dataset. NG is NDCG. CV is Coverage. Boldface denotes highest score.

Models	accuracy				diversity				novelty				repetitivene		
	HR@10	NG@10	HR@20	NG@20	CV@1	CV@5	CV@10	CV@20	CV@1	CV@5	CV@10	CV@20	R@5	R@10	R
GRU	0.2673	0.1878	0.3082	0.1981	0.2439	0.4695	0.5699	0.6632	0.1837	0.4139	0.5238	0.6267	14.25	29.44	6
GRU-SQN	0.2967	0.2094	0.3406	0.2205	0.2180	0.4114	0.4975	0.5763	0.1526	0.3489	0.4430	0.5299	14.62	30.19	6
GRU-SMORL	**0.3060**	**0.2103**	**0.3535**	**0.2224**	**0.2796**	**0.5369**	**0.6419**	**0.7353**	**0.2154**	**0.4871**	**0.6029**	**0.7064**	**13.53**	**28.02**	**5**
Caser	0.2302	0.1675	0.2628	0.1758	0.2327	0.4379	0.5133	0.5718	0.1643	0.3773	0.4605	0.5252	16.16	33.24	6
Caser-SQN	0.2454	0.1778	0.2803	0.1867	0.2088	0.3880	0.4511	0.5021	0.1387	0.3219	0.3914	0.4479	16.88	34.50	7
Caser-SMORL	**0.2657**	**0.1898**	**0.3052**	**0.1998**	**0.2855**	**0.5411**	**0.6324**	**0.7138**	**0.2224**	**0.4917**	**0.5925**	**0.6827**	**15.90**	**32.47**	**6**
NtItNet	0.3007	0.2060	0.3506	0.2186	0.2867	0.5113	0.6033	0.6837	0.2305	0.4595	0.5605	0.6495	12.25	25.76	5
NtItNet-SQN	0.3129	0.2150	0.3586	0.2266	0.2802	0.5255	0.6077	0.6750	0.2184	0.4747	0.5651	0.6395	12.27	25.93	5
NtItNet-SMORL	**0.3183**	**0.2222**	**0.3659**	**0.2342**	**0.3429**	**0.6335**	**0.7351**	**0.8129**	**0.2800**	**0.5938**	**0.7062**	**0.7924**	**10.92**	**22.89**	**4**
SASRec	0.3085	0.2107	0.3572	0.2227	0.2767	0.5305	0.6300	0.7149	0.2171	0.4806	0.5899	0.6838	15.67	32.27	6
SASRec-SQN	0.3302	0.2279	0.3803	0.2406	0.2393	0.4617	0.5490	0.6254	0.1753	0.4040	0.5001	0.5847	15.60	32.20	6
SASRec-SMORL	**0.3521**	**0.2477**	**0.4028**	**0.2605**	**0.3037**	**0.5724**	**0.6672**	**0.7476**	**0.2366**	**0.5261**	**0.6311**	**0.7202**	**12.58**	**26.69**	**5**

Woodstock '18, June 03–05, 2018, Woodstock, NY Dušan Stamenković, Alexandros Karatzoglou, Ioannis Arapakis, Xin Xin, and Kleomenis Katevas

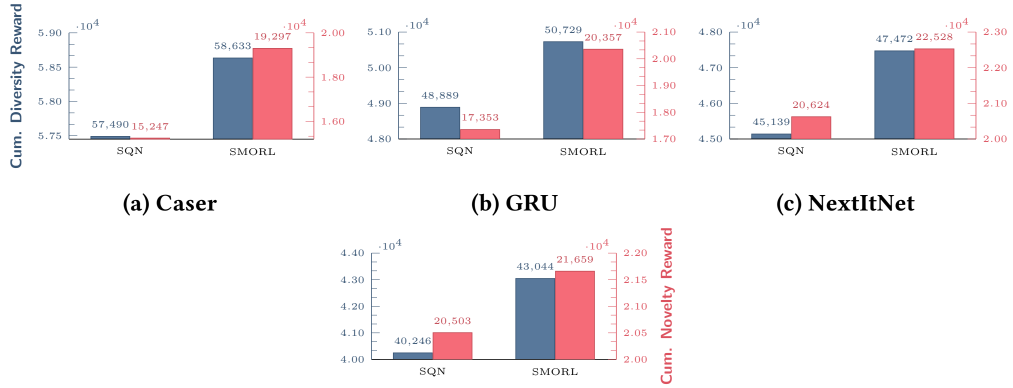
5.4 Gradient Intensity Investigation (RQ3)

在所有基本模型和两个数据集上，SDQL损失由自监督损失主导，这表明从公式(5)中优化参数α可能会改善SMORL部分对基本模型的影响。图4显示了NextItNet-SMORL模型在两个数据集上关于NDCG@20和CV@20指标的表现，当我们改变SDQL梯度的强度时。如预期，按$\alpha < 1$乘以SDQL时，效果降低，与基本模型相比我们没有显著改善。对于$\alpha \in \{1, 2, 3, 5, 10\}$，可以看到两个指标的增加，当$\alpha = 5$时获得了最佳平衡。对于更高的$\alpha$值，由于从自监督损失获得的梯度信号丧失，我们观察到质量显著下降，这表明需要有自监督部分来学习基本的排名。对于RC15数据集，也可以进行类似的分析。

在大多数情况下，α参数的最佳价值等于1 - 在RC15数据集上的SASRec，RetailRocket上的GRU4Rec，以及RetailRocket数据集上的Caser。然而，对于RC15上的GRU4Rec和Caser，最佳值等于0.75，对于RC15上的SASRec和NextItNet为3，而RetailRocket上的SASRec等于10。因此，对于真实世界的使用案例，数据集通常包含数百万项时，更高的α值可能是最佳的。更复杂的模型，如NextItNet和SASRec需要更高的α值。

6 CONCLUSIONS & FUTURE WORK

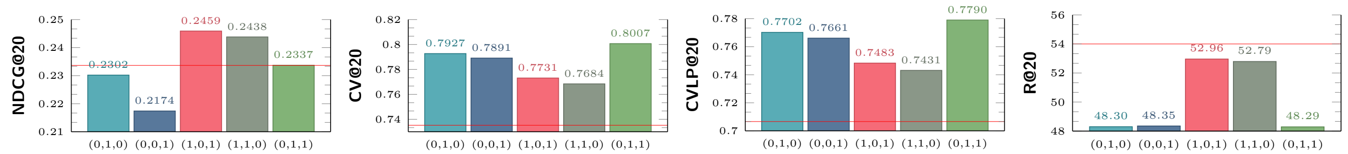
我们首先正式化了下一个项目推荐任务，并将其作为多目标MDP任务呈现。SMORL方法作为正则化器，用于将期望的特性引入推荐模型，特别是为了实现推荐的相关性、多样性和新颖性之间的平衡。我们将SMORL与四个最先进的推荐模型集成，并在两个真实的电子商务数据集上进行了实验。我们的实验发现，三项相互冲突目标的联合优化对于提高与用户满意度高度相关的指标至关重要，同时还保持内容的相关性。未来的工作为探索SMORL范式在推荐系统（RS）环境中的应用带来了广阔的可能性，并将包括与不同目标的进一步实验以及在不同领域（如音乐平台）中应用SMORL。此外，监督损失与SDQL损失的联合优化本身也是一个研究课题。最后，我们计划探索使用非线性和个性化的标量化函数。



(a) Caser (b) GRU (c) NextItNet

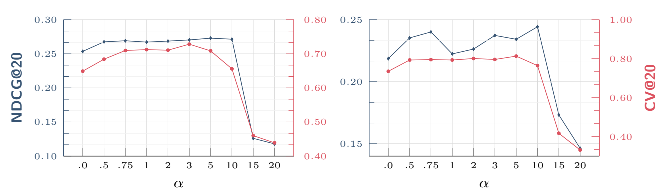
(d) SASRec

图 2: Comparison of cumulative rewards on RC15 dataset with base models regularized by SMORL and SQN frameworks.



(a) Accuracy (b) Diversity (c) Novelty (d) Repetitiveness

图 3: Performance comparison when reinforcing a subset of objectives (achieved by using different configurations of w from Eq.(7)). Red lines denote relevant metrics of the stock NextItNet model - not trained using the SMORL4RS framework.



(a) (b)

图 4: NextItNet with different intensity of SMORL gradient on the RC15 (4a) and RetailRocket (4b) datasets

challenge. In this section, our goal is to demonstrate that we can control how much influence each objective has, and not how to find an ideal balance. With the ability of control, many engineering possibilities arise, such as deploying multiple SMORL4RS agents and deciding in an online fashion if a user should receive recommendations from an agent that is optimized towards novelty, diversity, or accuracy.

Figure 3 shows the comparison of NextItNet-SMORL model regularized by the SMORL agent that uses mentioned weight configurations **w** on RetailRocket dataset, while similar behaviour can be observed for the RC15 dataset and other models. More specifically, Figure 3a indicates that if we regularize the model only towards novelty, we will sacrifice its ability to recommend relevant items. This phenomenon is also present if we only reinforce towards diversity, but the drop in NDCG@20 metric is not as notable. On the other hand, if we optimize jointly towards diversity and novelty, we do not observe a drop in the accuracy of the base model. Additionally, if we include the accuracy objective to any of the two other, we observe an increase in the relevant metric. From Figures 3b and 3c, we note that including the accuracy objective comes at the cost of diversity and novelty, while combined

REFERENCES

[1] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*. 2155–2165.

[2] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. 2015. Optimal Greedy Diversity for Recommendation.. In *IJCAI*, Vol. 15. 1742–1748.

[3] Aishwariya Budhrani, Akashkumar Patel, and Shivam Ribadiya. 2020. Music2Vec: Music Genre Classification and Recommendation System. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 1406–1411.

[4] Laming Chen, Guoxin Zhang, and Hanning Zhou. 2017. Fast greedy map inference for determinantal point process to improve recommendation diversity. *arXiv preprint arXiv:1709.05135* (2017).

[5] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 456–464.

[6] Xinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Yuan Qi, and Le Song. 2019. Generative adversarial user model for reinforcement learning based recommendation system. In *International Conference on Machine Learning*. PMLR, 1052–1061.

[7] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. 2017. Learning to recommend accurate and diverse items. In *Proceedings of the 26th international conference on World Wide Web*. 183–192.

[8] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[9] Daniel M Fleder and Kartik Hosanagar. 2007. Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on Electronic commerce*. 192–199.

[10] Christian Hansen, Rishabh Mehrotra, Casper Hansen, Brian Brost, Lucas Maystre, and Mounia Lalmas. 2021. Shifting Consumption towards Diverse Content on Music Streaming Platforms. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 238–246.

[11] Hado V Hasselt. 2010. Double Q-learning. In *Advances in neural information processing systems*. 2613–2621.

[12] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[13] 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 and knowledge management*. 843–852.

[14] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[15] Rong Hu and Pearl Pu. 2011. Helping Users Perceive Recommendation Diversity.

[16] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 368–377.

[17] Sheena S Iyengar and Mark R Lepper. 1999. Rethinking the value of choice: a cultural perspective on intrinsic motivation. *Journal of personality and social psychology* 76, 3 (1999), 349.

[18] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[19] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 210–217.

[22] Frédéric Lavancier, Jesper Møller, and Ege Rubak. 2015. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2015), 853–877.

[23] Ye Ma, Lu Zong, Yikang Yang, and Jionglong Su. 2019. News2vec: News network embedding with subnode information. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4845–4854.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[25] Nikola Milojkovic, Diego Antognini, Giancarlo Bergamin, Boi Faltings, and Claudiu Musat. 2019. Multi-gradient descent for multi-objective recommender systems. *arXiv preprint arXiv:2001.00846* (2019).

[26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[27] Hossam Mossalam, Yannis M Assael, Diederik M Roijers, and Shimon Whiteson. 2016. Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707* (2016).

[28] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.

[29] Lijing Qin and Xiaoyan Zhu. 2013. Promoting diversity in recommendation by entropy regularizer. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer.

[30] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[31] Marco Tulio Ribeiro, Anisio Lacerda, Adriano Veloso, and Nivio Ziviani. 2012. Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the sixth ACM conference on Recommender systems*. 19–26.

[32] Barry Schwartz. 2004. The paradox of choice: Why less is more. *New York: Ecco* (2004).

[33] Chaofeng Sha, Xiaowei Wu, and Junyu Niu. 2016. A Framework for Recommending Relevant and Diverse Items.. In *IJCAI*, Vol. 16. 3868–3874.

optimization towards diversity and novelty produce the best results with respect to these metrics. Similarly, by including the accuracy objective, we increase the repetitiveness compared to the NextItNet-SMORL model that optimizes towards a combination of diversity and novelty.

5.4 Gradient Intensity Investigation (RQ3)

Across all base models and both datasets, the SDQL loss is dominated by self-supervised loss, which suggests that the optimization of parameter α from Eq.(5) might improve the effect of SMORL part on the base model. Figure 4 shows the behaviour of NextItNet-SMORL model with respect to NDCG@20 and CV@20 metrics on both datasets when we change the intensity of SDQL gradient. As expected, when multiplying SDQL with $\alpha < 1$, the effects are decreased and we do not improve dramatically compared to the base model. Increase in both metrics can be seen for $\alpha \in \{1, 2, 3, 5, 10\}$, with the best balance obtained for $\alpha = 5$. For higher values of α, we observe a notable drop in quality due to the loss of gradient signal obtained from the self-supervised loss, which indicates that it is necessary to have a self-supervised part to learn basic ranking. Similar analysis can be made for RC15 dataset.

The optimal value of the α parameter is equal to 1 for most cases - SASRec on RC15 dataset, GRU4Rec on RetailRocket, and Caser on the RetailRocket dataset. However, for GRU4Rec and Caser on RC15, the optimal value is equal to 0.75, for SASRec and NextItNet on RC15 to 3, while for SASRec on RetailRocket is equal to 10. Hence for real-world use-cases, when datasets usually contain millions of items, higher values of α might be optimal. More complex models, such as NextItNet and SASRec require higher value of α.

6 CONCLUSIONS & FUTURE WORK

We first formalized the next item recommendation task and presented it as a Multi-Objective MDP task. The SMORL method acts as a regularizer for introducing desirable properties into the recommendation model, specifically to achieve a balance between relevance, diversity and novelty of recommendations. We integrated SMORL with four state-of-the-art recommendation models and conducted experiments on two real-world e-commerce datasets. Our experimental findings demonstrate that the joint optimization of three conflicting objectives is essential for improving metrics that are strongly correlated with user satisfaction, while also preserving content relevance. Future work brings vast possibilities for exploring the use of SMORL paradigm in the setting of RS, and it will include further experiments with different objectives and application of SMORL in different areas, such as music platforms. Also, the joint optimization of supervised and SDQL loss is a research problem on its own. Finally, we plan on exploring the use of non-linear and personalized scalarization functions.

[34] Wenjie Shang, Yang Yu, Qingyang Li, Zhiwei Qin, Yiping Meng, and Jieping Ye. 2019. Environment reconstruction with hidden confounders for reinforcement learning based recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 566–576.

[35] Cass R Sunstein and Edna Ullmann-Margalit. 1999. Second-order decisions. *Ethics* 110, 1 (1999), 5–31.

[36] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 565–573.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[38] Romain Warlop, Jérémie Mary, and Mike Gartrell. 2019. Tensorized determinantal point processes for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1605–1615.

[39] C Ch White, CC III WHITE, and KIM KW. 1980. Solution procedures for vector criterion Markov decision processes. (1980).

[40] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H Chi, and Jennifer Gillenwater. 2018. Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2165–2173.

[41] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[42] Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M Jose. 2020. Self-Supervised Reinforcement Learning forRecommender Systems. *arXiv preprint arXiv:2006.05779* (2020).

[43] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 582–590.

[44] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 13–22.

[45] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. 2018. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1040–1048.

[46] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 167–176.

[47] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. 2019. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2810–2818.

REFERENCES

[1] Ashton Anderson, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. Algorithmic effects on the diversity of consumption on spotify. In *Proceedings of The Web Conference 2020*. 2155–2165.

[2] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. 2015. Optimal Greedy Diversity for Recommendation.. In *IJCAI*, Vol. 15. 1742–1748.

[3] Aishwariya Budhrani, Akashkumar Patel, and Shivam Ribadiya. 2020. Music2Vec: Music Genre Classification and Recommendation System. In *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. IEEE, 1406–1411.

[4] Laming Chen, Guoxin Zhang, and Hanning Zhou. 2017. Fast greedy map inference for determinantal point process to improve recommendation diversity. *arXiv preprint arXiv:1709.05135* (2017).

[5] Minmin Chen, Alex Beutel, Paul Covington, Sagar Jain, Francois Belletti, and Ed H Chi. 2019. Top-k off-policy correction for a REINFORCE recommender system. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 456–464.

[6] Xinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Yuan Qi, and Le Song. 2019. Generative adversarial user model for reinforcement learning based recommendation system. In *International Conference on Machine Learning*. PMLR, 1052–1061.

[7] Peizhe Cheng, Shuaiqiang Wang, Jun Ma, Jiankai Sun, and Hui Xiong. 2017. Learning to recommend accurate and diverse items. In *Proceedings of the 26th international conference on World Wide Web*. 183–192.

[8] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[9] Daniel M Fleder and Kartik Hosanagar. 2007. Recommender systems and their impact on sales diversity. In *Proceedings of the 8th ACM conference on Electronic commerce*. 192–199.

[10] Christian Hansen, Rishabh Mehrotra, Casper Hansen, Brian Brost, Lucas Maystre, and Mounia Lalmas. 2021. Shifting Consumption towards Diverse Content on Music Streaming Platforms. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 238–246.

[11] Hado V Hasselt. 2010. Double Q-learning. In *Advances in neural information processing systems*. 2613–2621.

[12] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[13] 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 and knowledge management*. 843–852.

[14] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2015. Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939* (2015).

[15] Rong Hu and Pearl Pu. 2011. Helping Users Perceive Recommendation Diversity.

[16] Yujing Hu, Qing Da, Anxiang Zeng, Yang Yu, and Yinghui Xu. 2018. Reinforcement learning to rank in e-commerce search engine: Formalization, analysis, and application. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 368–377.

[17] Sheena S Iyengar and Mark R Lepper. 1999. Rethinking the value of choice: a cultural perspective on intrinsic motivation. *Journal of personality and social psychology* 76, 3 (1999), 349.

[18] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)* 20, 4 (2002), 422–446.

[19] Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 197–206.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Neal Lathia, Stephen Hailes, Licia Capra, and Xavier Amatriain. 2010. Temporal diversity in recommender systems. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 210–217.

[22] Frédéric Lavancier, Jesper Møller, and Ege Rubak. 2015. Determinantal point process models and statistical inference. *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2015), 853–877.

[23] Ye Ma, Lu Zong, Yikang Yang, and Jionglong Su. 2019. News2vec: News network embedding with subnode information. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4845–4854.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[25] Nikola Milojkovic, Diego Antognini, Giancarlo Bergamin, Boi Faltings, and Claudiu Musat. 2019. Multi-gradient descent for multi-objective recommender systems. *arXiv preprint arXiv:2001.00846* (2019).

[26] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[27] Hossam Mossalam, Yannis M Assael, Diederik M Roijers, and Shimon Whiteson. 2016. Multi-objective deep reinforcement learning. *arXiv preprint arXiv:1610.02707* (2016).

[28] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[29] Lijing Qin and Xiaoyan Zhu. 2013. Promoting diversity in recommendation by entropy regularizer. In *Twenty-Third International Joint Conference on Artificial Intelligence*. Citeseer.

[30] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[31] Marco Tulio Ribeiro, Anisio Lacerda, Adriano Veloso, and Nivio Ziviani. 2012. Pareto-efficient hybridization for multi-objective recommender systems. In *Proceedings of the sixth ACM conference on Recommender systems*. 19–26.

[32] Barry Schwartz. 2004. The paradox of choice: Why less is more. *New York: Ecco* (2004).

[33] Chaofeng Sha, Xiaowei Wu, and Junyu Niu. 2016. A Framework for Recommending Relevant and Diverse Items.. In *IJCAI*, Vol. 16. 3868–3874.

[34] Wenjie Shang, Yang Yu, Qingyang Li, Zhiwei Qin, Yiping Meng, and Jieping Ye. 2019. Environment reconstruction with hidden confounders for reinforcement learning based recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 566–576.

[35] Cass R Sunstein and Edna Ullmann-Margalit. 1999. Second-order decisions. *Ethics* 110, 1 (1999), 5–31.

[36] Jiaxi Tang and Ke Wang. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. 565–573.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[38] Romain Warlop, Jérémie Mary, and Mike Gartrell. 2019. Tensorized determinantal point processes for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1605–1615.

[39] C Ch White, CC III WHITE, and KIM KW. 1980. Solution procedures for vector criterion Markov decision processes. (1980).

[40] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H Chi, and Jennifer Gillenwater. 2018. Practical diversified recommendations on youtube with determinantal point processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2165–2173.

[41] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.

[42] Xin Xin, Alexandros Karatzoglou, Ioannis Arapakis, and Joemon M Jose. 2020. Self-Supervised Reinforcement Learning forRecommender Systems. *arXiv preprint arXiv:2006.05779* (2020).

[43] Fajie Yuan, Alexandros Karatzoglou, Ioannis Arapakis, Joemon M Jose, and Xiangnan He. 2019. A simple convolutional generative network for next item recommendation. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 582–590.

[44] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 13–22.

[45] Xiangyu Zhao, Liang Zhang, Zhuoye Ding, Long Xia, Jiliang Tang, and Dawei Yin. 2018. Recommendations with negative feedback via pairwise deep reinforcement learning. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1040–1048.

[46] Guanjie Zheng, Fuzheng Zhang, Zihan Zheng, Yang Xiang, Nicholas Jing Yuan, Xing Xie, and Zhenhui Li. 2018. DRN: A deep reinforcement learning framework for news recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 167–176.

[47] Lixin Zou, Long Xia, Zhuoye Ding, Jiaxing Song, Weidong Liu, and Dawei Yin. 2019. Reinforcement learning to optimize long-term user engagement in recommender systems. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2810–2818.