基于强化学习的推荐的离线评估：一个关键问题及若干替代方案

Romain DeffayetNaver Labs Europe &阿姆斯特丹大学 法国 / 荷兰
romain.deffayet@naverlabs.com

Thibaut ThonetNaver Labs Europe 法国
thibaut.thonet@naverlabs.com

Jean-Michel Renders
Naver Labs Europe
法国
jean-michel.renders@naverlabs.com

Maartende Rijke阿姆斯特丹大学
荷兰m.derijke@uva.nl

arXiv:2301.00993v1 [cs.IR] 3 Jan 2023

摘要

本文认为，通常用于离线评估序列推荐系统的范式并不适合评估基于强化学习的推荐系统。我们发现，现有的基于强化学习的推荐的离线评估实践大多采用下一物品预测 协议，并详细阐述了这种评估协议的三个缺陷。值得注意的是，它无法体现强化学习（RL）预期带来的潜在益处，却掩盖了某些离线强化学习智能体的关键缺陷。我们提出了评估基于强化学习的推荐系统的替代方法，旨在阐明现有的可行方案，并启发未来关于可靠评估协议的研究。

1 引言

推荐系统无处不在，例如广泛应用于内容服务平台和电子商务平台，因此在塑造互联网用户的体验方面发挥着重要作用。因此，正确、审慎地评估推荐系统至关重要，因为这会直接影响业务指标和用户满意度，有时甚至影响整个社会。

虽然推荐准确率（即推荐相关项目）通常被视为衡量性能的主要指标，但推荐系统领域的文献强调了其他标准的重要性。超越准确率的目标包括多样性、新颖性或意外惊喜、公平性，以及总体上的用户体验 [McNee 等人， 2006]。此类标准有时无法通过单次推荐（即用户与推荐系统之间的一次交互）来实现，而可能需要我们考量更长期的体验。这些问题促使研究人员和从业者都认识到许多推荐引擎的序列性

，并寻求针对完整序列而非单次预测进行优化 [Quadrana等人， 2018]。

强化学习（RL）将这一问题建模为马尔可夫决策过程（MDP），在该过程中，我们希望选择合适的动作（即物品推荐），以最大化用户与推荐系统进行完整交互序列中的奖励总和（例如点击、购买等）。强化学习非常适合这一问题，因为底层的 MDP 能够建模推荐对用户产生的长期影响。请注意，在推荐场景中，通常无法进行在线探索，因此必须基于固定数据集中的交互来训练策略，也就是采用离线强化学习。虽然离线强化学习的序列优化未必能完全满足文献中强调的所有超越准确率的标准，但它有望通过对完整序列进行优化，自然而然地实现一些期望的特性。事实上，可以预期，只要奖励函数设置得当，能够贯穿用户整个体验过程并发挥作用的策略就需要具备一些期望的特性：多样性、新颖性等。由于这些辅助指标已融入序列的累积奖励中，因此，使用强化学习对完整序列进行优化，可以视为弥合离线与在线性能差距的一种方式。

本文认为，借助强化学习，序列推荐所取得的进展看似显著，却缺乏生态效度 [Andrade, 2018]：由于当前评估实践存在一些不足，训练好的智能体很可能无法泛化到现实场景。具体而言，基于强化学习的推荐系统通常采用离线方式进行评估，沿用传统的单次准确率导向协议，而这种协议无法捕捉使用强化学习算法所带来的潜在益处。我们将这种评估协议称为下一物品预测（NIP）。更值得关注的是，我们指出，这一协议的具体设计很可能掩盖通过离线强化学习训练的推荐系统存在的缺陷。简言之，我们认为，按照最常用的评估实践，我们无法验证强化学习算法是否正确地优化了其旨在优化的指标，即期望累积奖励。我们担心，这种做法非但没有缩小离线性能与在线性能之间的差距，反而会使差距进一步扩大。随后，我们针对基于强化学习的推荐提出合理的评估方法建议，以帮助从业者和研究人员避免常见陷阱，并启发这一重要领域的未来研究。

在第2节将我们的批评与先前研究提出的批评进行对比后，我们会在第3 节定义下一物品预测评估协议，并概述该协议在基于强化学习的序列推荐中的应用。第4 节深入探讨NIP协议的三个主要问题，以及这些问题对基于强化学习的推荐系统评估的影响。最后，我们将在第5节提出关于如何在基于强化学习的推荐中采用合理评估方法的建议。

2 相关研究

推荐系统评估的缺陷一直是推荐领域长期存在的问题。本节将回顾讨论这一主题的先前研究。

首先，正如我们在引言中所回顾的，McNee 等人 [2006]；Jannach 等人 [2016] 强调了开发超越所推荐物品准确率的推荐系统的必要性，也就是说，不应仅仅将推荐视为矩阵补全问题。这一观点源于

人们观察到离线性能与在线性能之间存在差距，这有时会使基于离线评估得出的任何结论都失去意义 [Garcin 等人，2014年；Gomez-Uribe 和 Hunt，2016年；Jeunen，2019]。

其次，推荐系统评估中的陷阱——包括本研究重点关注的离线评估下一物品预测协议——过去已有大量讨论：Chen 等人 [2017]；Jeunen [2019]；Ji 等人 [2020]；Cremonesi 和 Jannach [2021]；Sun[2022]；Zhao 等人 [2022] 指出，数据泄漏及其他数据集构建谬误会引发多种问题，并可能导致违反直觉的结论。研究者还发现，用于评估基于隐式反馈的推荐系统的数据中存在选择偏差，这是造成评估结果不准确的一个主要原因 [Gomez-Uribe 和 Hunt，2016年；Jannach 等人，2016年；Chen 等人，2017年；Jeunen， 2019]。此外，针对下一物品预测协议，Krichene 和 Rendle [2020]；Zhao 等人 [2022] 表明，为简化排序指标的计算而在推理阶段对负样本物品进行采样，会导致对推荐性能得出错误结论。

最后，许多研究再次证实，适当选择基线方法对于确保取得进展至关重要，并表明某些主张在经过恰当调优的基线方法面前并不成立 [Ludewig 等人，2019年；Ferrari Dacrema 等人，2019年；Rendle 等人，2019年；Sun 等人，2020年；Zhao 等人， 2022]。

本文提出的论点专门针对基于强化学习的推荐。尽管据我们所知，这一论点此前从未被明确提出，但它并不与本节所列的问题相矛盾。评估基于强化学习的推荐系统时，应将其视为一项额外的注意事项。

3 基于强化学习的推荐中的下一物品预测

我们提出了一个（非正式的）下一物品预测定义，涵盖了许多序列推荐研究所采用的离线评估协议；我们认为，将这些协议用于评估基于强化学习的推荐系统存在问题：

定义1。下一物品预测（NIP）是一种基于真实用户反馈、用于序列物品推荐的离线评估协议。该任务旨在确保根据历史交互序列，用户下一次交互的物品位列模型排名靠前的物品之中。模型性能根据排序指标（例如命中率、召回率、归一化折损累计增益等）进行衡量。

我们提出这一定义，是因为它代表了许多序列推荐研究所采用的评估设置，例如 GRU4REC [Hidasi 等人， 2016], 也涵盖了若干变体。具体而言，“下一个交互物品”的选择会因数据集和具体任务而异：内容推荐中的下一个点击物品（例如 Last.fm [Last.fm]）、产品推荐中的下一个购买的商品（例如 RecSys Challenge 2015 [Ben-Shimon 等人， 2015] 或 RetailRocket [RetailRocket， 2016]）、电影推荐中的下一个高评分电影（例如 MovieLens [GroupLens]）、食品杂货购物中的下一个购物篮 [Instacart， 2017],等等。

它在基于强化学习的推荐中有多普遍？ 近年来，基于强化学习的推荐（RL4REC）越来越受欢迎：我们统计了2017年1月至2022年10月期间主要信息检索和推荐系统（或相关）会议论文集中关于RL4REC的55篇论文，

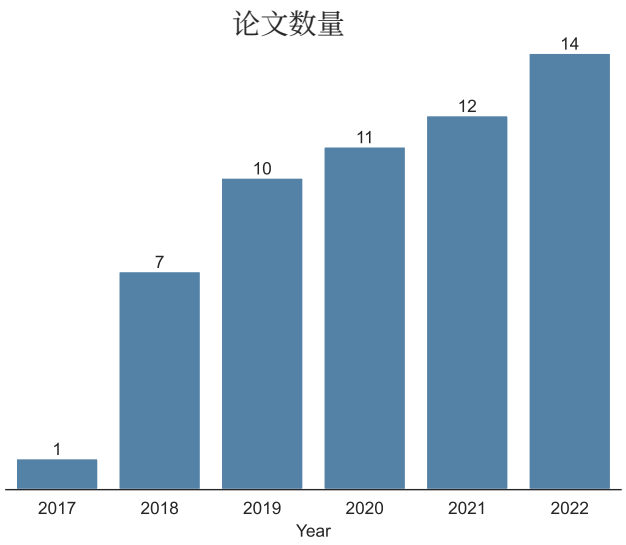


图 1：2017年到2022年间，在主要 RecSys 和 IR 会议上发表的基于强化学习的推荐论文数量变化。

为获得这一结果，我们在DBLP上检索了“reinforcement learning recommendation”和“reinforcement learning recommender”，[1] 并纳入了发表于AAAI、CIKM、ICDM、IJCAI、KDD、RecSys、SIGIR、WSDM或WWW的论文。图1展示了已发表RL4REC论文数量的增长趋势。在从DBLP检索到的55篇论文中，我们确定了39篇使用基于强化学习的方法处理序列物品推荐的论文。与我们的论点无关的其他任务包括对话式推荐或可解释推荐，因此本研究不考虑与这些主题相关的论文。在这39篇相关论文中，我们发现24篇进行了某种形式的离线评估，其中22篇遵循了定义1中的NIP协议。另外15篇论文完全依赖在线评估，要么在生产环境中使用工业推荐平台进行评估，要么基于模拟器进行评估。因此，*NIP*协议是迄今为止最常采用的离线评估类型。

4 下一物品预测的三个缺点

在解释下一物品预测存在的问题之前，我们先回顾一下使用强化学习算法所带来的优势：

- 强化学习旨在优化由决策序列带来的长期结果。这需要将推荐系统对用户产生的影响纳入考量。基于强化学习的方法能够通过将观测到的奖励归因于单个动作来优化完整序列，从而避免在动作序列的组合空间中进行代价高昂的搜索。

[1] https://dblp.org/

- 强化学习算法通过最大化标量奖励，以自监督的方式进行学习。这样便能找到开放式解决方案，并生成新颖的策略。然而，以离线方式训练智能体也存在风险：所得策略的期望回报估计可能不准确。

下面我们列举用于评估离线强化学习智能体的 NIP 协议的三个主要缺陷，并说明这些缺陷如何损害基于该评估协议得出的结论的生态效度。

4.1 短视的评估

根据定义评估基于离线强化学习的推荐系统1 仅考虑短期奖励，却忽略了推荐对用户产生的因果效应。事实上，设计强化学习算法的一个重要动机，是最大化完整轨迹上的回报（即奖励之和）；这与旨在最大化每个时间步平均奖励的老虎机算法不同。当动作（即推荐）导致环境（即用户）改变其状态时，强化学习算法仍具有收敛保证，而对于老虎机算法而言，环境则呈现非平稳性，因此这类算法无论在理论上还是实践中都无法找到最优策略。然而，下一物品预测评估协议只要求进行短期思考，因为它奖励对下一个交互物品的单次预测——这是离线评估的静态特性所致，该特性忽视了所关注的推荐策略对后续交互产生的因果影响。 Lee等人 [2022], 对此进行了阐述，并通过实证验证，在标准推荐数据集上，贪心的短视智能体在 NIP 协议上的表现与具备长期意识的强化学习智能体相近，甚至更好。 Quadrana等人[2018] 也指出，在序列推荐中，如果目标不仅包括即时满意度，还包括多样性或在内容发现过程中为用户提供引导，那么 NIP 评估协议存在局限。

然而，与Lee等人 [2022], 不同，我们还认为，在内容推荐中纳入停留时间等延迟奖励，或在产品推荐中纳入生命周期价值，并不足以解决这一问题。事实上，延迟奖励所编码的长期结果（例如，产品在整个生命周期内是否令人满意？）可能与沿轨迹的奖励总和所编码的长期结果（例如，这条轨迹是否足够多样，以免让用户感到厌倦？）毫不相关。前者显然有助于构建实用且令人愉悦的推荐系统，而后者则是强化学习智能体所依赖的马尔可夫决策过程中建模的对象。因此，如果我们纳入延迟奖励，却忽略序列决策过程所带来的长期结果，仍然无法通过NIP协议观察到强化学习训练带来的益处。请注意，这两类长期结果并不矛盾；我们建议采用尽可能贴合用户需求与满意度的奖励函数，并将延迟结果纳入考量。

4.2 一个次优目标

如第3节所述，在常用于下一物品预测的数据集中，我们只能观察到用户与之交互的物品所获得的奖励（例如点击、购买）。这会在评估协议中引入选择偏差，其原因是对

用户施加了特定的处理。这种处理可以是日志策略，或者是在推荐平台上收集自然交互数据时采用的多种日志策略的组合；也可以是外生因素产生的隐式影响，前提是观测数据源于用户主动反馈，例如自愿撰写的电影评论或商品搜索。为简便起见，我们将后一类偏差称为隐式日志策略。需要注意的是，用户交互物品并非最优的另一个原因是，用户选择可能目光短浅，或不愿接受新颖性，即使这样做可能导致整体体验不那么愉快。

考虑到选择用户交互过的物品是一个二元目标，而非需要最大化的标量奖励，下一项预测评估会激励研究人员和从业者构建接近日志策略（隐式）的策略，代价则是无法选择最优动作。这是一项封闭式的策略匹配任务，而强化学习允许开放式结果，即生成能够获得高回报的新策略。存在更简单的方法来复现生成数据的策略，例如模仿学习 [Hussein 等人， 2017], 而强化学习的奖励最大化目标很可能会因选择不同于用户交互物品、但能带来更高回报的物品而降低这项评估的结果。因此，下一物品预测会舍弃表现出色的策略，并鼓励采用类似日志策略的策略，即使数据集中的序列极不理想。考虑购买或高评分等更强的信号可以缓解这一问题，但用户在数据收集期间接触到哪些物品存在选择偏差，这意味着一些奖励极高的物品很可能会被舍弃。

4.3 高风险部署

我们提出的前两个要点表明，下一物品预测评估无法体现基于离线强化学习的推荐系统所带来的潜在优势。我们讨论的第三个问题则表明，下一物品预测也可能掩盖离线强化学习智能体的关键缺陷。

尽管在定义 1的评估协议中，我们通过使用排序指标，将下一次交互物品在模型预测中的位置纳入考量，但推荐系统在生产环境中使用时，只会选择其最偏好的物品（或列表推荐中排名前k 位的最偏好物品），而不会向用户展示其他任何物品。因此，无论完整排序结果如何，确保排名第一的物品能够满足用户需求似乎至关重要。然而，对于固定数据集而言，这一点遗憾地无法实现，因为其中只向所考察的用户展示过一个或少数几个物品。下一物品预测的一个隐含假设是，更高的排序指标与能够带来高回报的排名第一物品相关。然而，先前研究已发现离线与在线结果之间存在差距 [Garcin 等人，2014年；Gomez-Uribe 和 Hunt， 2016]。更重要的是，研究表明，即使在强假设下，即每个动作（即物品推荐）所对应的 Q价值能够在期望意义上被正确估计（即不存在偏差），由于所选物品更有可能是那些被高估其 Q价值 [Jeunen 和 Goethals， 2021]，预测的离线奖励仍可能高于实际在线奖励。这种现象称为优化器诅咒。尽管在某些情况下其实际影响可能有限，但我们认为它会严重影响强化学习算法。事实上，人们已经发现，一组特定条件会导致优化器诅咒产生灾难性影响，这通常被称为致命三元组 [van Hasselt 等人，2018年；Sutton 和 Barto， 2018]。大多数强化学习算法都可能出现这种现象，其发生条件为： （i）使用一个状态的价值估计

来更新前一状态的价值估计；（ii）使用函数逼近来构建价值函数的估计；以及（iii）强化学习智能体以离策略方式进行训练。

在这种情况下，分布外动作上的价值函数即使只有轻微的高估，也可能被放大并传播到相邻的状态和动作，最终可能导致价值函数发散。在这种情形下，尽管模型为其策略预测了很高的 Q值，但部署后观察到的回报可能差得离谱。致命三元组造成的严重危害已在多个场景中得到观察，并推动了离线强化学习领域大量研究的兴起 [van Hasselt 等人，2018年；Fu 等人，2019年、2020年；Levine 等人，2020年；Brandfonbrener 等人，2021年；Kostrikov 等人，2021]。遗憾的是，在定义*1*中的标准下一物品预测评估中，无法检测到这种有害现象：模型可能会合理地将用户交互过的物品排在较高位置，但很可能至少有一个分布外物品被模型大幅高估并优先选择。由于模型会选择该物品，部署时可能会出现无法预料的灾难性故障。更糟的是，随着动作空间规模的扩大，故障概率往往也会增加 [Gu 等人， 2022], 而在某些推荐场景中，动作空间可能极其庞大。

4.4 结论

本节指出的三个缺陷，使得使用强化学习的基于下一物品预测协议评估基于强化学习的推荐系统并不可靠。这实际上扩大了离线指标与在线指标之间的差距，而强化学习算法本应弥合这一差距。下一节中，我们将提出解决这一问题的潜在方案。

5 下一物品预测的若干替代方案

下一物品预测的局限性使得基于强化学习的推荐系统难以进行离线评估。下面我们将详细介绍一些针对这一问题的部分解决方案，并讨论其局限性以及尚待解答的问题。

5.1 推荐平台中的在线评估

针对上述问题，最显而易见的应对措施是在可能的情况下，直接根据我们关注的指标对推荐系统进行在线评估。这通常通过在实际的推荐平台上部署策略来实现。然而，显然并非所有研究人员和从业者都能使用正在运行的工业平台，而且在线评估本身也可能引入其他形式的偏差，例如在推荐中纳入业务规则。在线评估显然能够规避上一节强调的三个问题，但由于本文聚焦于离线评估，因此不再对此作进一步详述。

5.2 反事实离策略评估

信息检索领域关于离策略评估（OPE）的研究成果十分丰富，通常基于逆倾向评分等技术 [Swaminathanand Joachims，2015年；Joachims 等人， 2017], 其中，通过应用倾向权重对观测到的奖励和回报进行重新缩放。

尽管离策略评估主要针对单步老虎机问题展开，但一些研究分别在强化学习社区 [Fu et al., 2021] 和信息检索社区 [Chen 等人， 2019], 探讨了强化学习策略的离策略评估；最近， [Kiyohara and Kawakami 提出了一个用于信息检索领域强化学习算法离策略评估的库 2022]。

反事实方法用于离策略评估颇具吸引力，因为在温和的假设下，它们能够保证无偏性。然而，我们想强调这类方法存在三个（已知的）缺陷：(i) 逆倾向评分的方差高得臭名昭著；应用于序列时，由于逆倾向权重的乘积，方差会呈指数级增大 [Precup 等人， 2000]；(ii) 在非表格设置中（即可以将对一个状态-动作对的预测推广到另一个状态-动作对的情形，例如连续空间），当日志策略未知时，必须隐式或显式地假设模型具有泛化能力，才能计算倾向得分 [Hanna 等人， 2019]；(iii) 当我们以离线方式训练强化学习算法时，离策略训练误差和离策略评估误差很可能相关，这意味着反事实离策略评估仍可能存在偏差，并错误地认为某些方法优于其他方法。后一种情况的一个极端例子是：如果我们使用相同的倾向权重训练并评估一个策略梯度推荐系统，那么无论其真实性能如何，智能体看起来都会是最优的。使用估计器集成或许能缓解这一问题，但如何彻底消除该问题仍不明确。从理论上讲，反事实离策略评估避免了上一节强调的所有三个缺陷，但正如我们所见，它自身也存在一些缺陷，因此在某些实际场景中可能并不可靠。

5.3 基于模拟器的评估

模拟器已证明在其他领域也很有用，例如机器人学、游戏或工业应用 [Fu 等人，2020年；Gulcehre 等人，2020年；Qin 等人， 2021]。由于人类行为的复杂性和明显的随机性，与推荐系统的交互无疑是最难模拟的问题之一，但模拟器的真正价值在于，它们能够观察推荐系统在一组既定的用户行为假设下如何响应。此外，通过让研究人员访问原本无法观察到的指标，模拟器还能帮助我们了解所构建系统的内部运作机制。

许多研究提出构建半合成模拟器，其中合成部分应尽可能有限，以贴近现实场景。例如，可以使用真实物品嵌入 [Shi 等人， 2019] 也可以将隐式反馈扩展到未见数据，并针对非随机缺失情形进行去偏 [Huang 等人， 2020]。此外，通过在广泛的模拟配置上进行基准测试，可以评估一种方法的可泛化性，从而减轻模拟器设计对结果的影响。无论选择何种设置，都应确保模拟器具备我们希望建模的特征，其中最值得关注的是推荐系统对用户产生的长期影响。

模拟器不受 NIP 协议这三个问题的影响，但其生态效度显然可能有限。除了基于真实数据构建模拟器外，一些方法还旨在弥合仿真与现实之间的差距，例如采用领域随机化 [Tobin等人，2017年；OpenAI等人， 2020]。

5.4 中间评估

我们所说的中间评估，是指对最终推荐模型中用作构建模块的模型、模拟器或倾向性进行离线评估 [Huang等人，2020年；Deffayet等人，2022]。在某些情况下，评估这些中间模型可能比评估最终模型更容易，例如，当有人工标注（如物品相关性标注）可供使用时。将评估协议拆分为若干组成部分，可以帮助我们隔离并减少偏差来源。例如，在以累计点击最大化为目标的前k项推荐中，如果点击模型估计准确，即相关性和倾向性分数都正确，那么剩下的不确定性来源就只有状态动态（即用户如何因推荐而改变）。

这样做可以降低部署强化学习智能体所带来的风险，但无法完全消除这些风险。此外，我们想强调，离线强化学习智能体很可能会在训练分布之外使用中间模型来进行策略评估，因此，如果没有采取适当的应对措施，它们可能会利用这些高不确定性区域中的不准确之处 [Deffayet 等人，2022]。

5.5 不确定性感知评估

虽然纯离线方式可能无法准确评估强化学习策略的最终表现，但我们认为，量化其在不同不确定性水平下的表现，有助于评估部署风险。事实上，上一节强调的价值高估问题，源于分布外状态-动作对存在较高的不确定性。我们可以约束强化学习算法，使其恢复出安全策略，让策略始终处于日志策略的分布范围内；也可以允许探索，以寻找潜在的高回报策略，但代价是不确定性会增加 [Brandfonbrener 等人，2021]。通过量化日志策略与目标策略的支持集之间的匹配程度，我们可以评估部署目标策略所带来的风险。具体而言，如果将可选动作限制为被视为“支持集内”的动作，就可以准确估计策略在这些动作上的表现。事实上，日志策略的支持集内不确定性较低；此外，还可以像 Q等研究所做的那样，在离线数据集的留出测试集上评估 [Ji 等人，2021] 对 Q值预测的质量。一项支持集内期望回报较高的安全策略，将构成可靠的改进；而一项连良好的支持集内期望回报都无法达到的不安全策略，则很可能可以舍弃。这类评估需要对支持集内和支持集外作出恰当定义，例如 [Fujimoto 等人，2019年；Gu 等人，2022],这在非表格设定中并非易事，而且需要假设对不确定性有一定的容忍度。不过，Kumar 等人 [2021] 表明，可以根据某些离线强化学习算法的训练曲线调整这一容忍度。

这类评估侧重于描述并缓解我们在第 4.3 节提出的第三个问题所引发的风险，同时或许也能让我们发现强化学习训练带来的益处。一个尚待解决的主要问题是，如何恰当地定义日志策略与目标策略支持集之间的距离度量。

6 结论

本研究指出，目前最常用于离线评估基于强化学习的推荐系统的方案实际上并不适用，因为它既无法体现强化学习相较于更传统方法所带来的益处，也可能掩盖离线强化学习智能体的严重缺陷，进而导致灾难性的部署后果。这些不足可概括如下：(i) 只衡量短期准确率的短视方案；(ii) 封闭且次优的推荐目标；以及 (iii) 对优化器诅咒的敏感性。

目前，对于如何完全以离线方式评估强化学习策略，尚无真正令人满意的解决方案。不过，我们可以使用若干在线性能的替代指标，弥合离线指标与在线性能之间的差距。在离线强化学习文献中，寻找合适的离线评估协议仍是一个活跃的研究领域，我们呼吁序列推荐领域的研究者加入这一工作，为推荐场景制定合适的协议。此外，承认基于强化学习的推荐系统在部署时存在不确定性，将有助于找到能够抵御或适应此类不确定性的解决方案。例如，Oosterhuis 和 de Rijke [2021] 提出了一种在遇到分布外情况时回退到更安全策略的准则（尽管其研究背景不同，即反事实学习排序），而 Ghosh 等人 [2022]；Reichlin 等人 [2022] 则提出了自适应离线强化学习策略，这些策略能够在部署过程中进入不确定状态后，通过分支回到有支持的状态来恢复。我们希望未来推荐系统领域的研究能更加重视这些方面，并缩小离线性能与在线性能之间的差距。

参考文献

Chittaranjan Andrade。研究设计、实施与评估中的内部效度、外部效度和生态效度。《印度心理医学杂志》，40:498–499，2018年。

David Ben-Shimon、Michael Friedmann、Alexander Tsikinovsky、Johannes H¨orle、Lior Rokach 和 Bracha Shapira。RecSys 挑战赛 2015，2015年。网址：https://recsys.acm.org/recsys15/challenge/。

David Brandfonbrener、Will Whitney、Rajesh Ranganath 和 Joan Bruna。无需离策略评估的离线强化学习。载于*NeurIPS*，第 4933–4946 页，2021年。Hung-Hsuan Chen、Chu-An Chung、Hsin-Chien Huang 和 Wen Tsui。推荐系统训练与评估中的常见陷阱。*ACM SIGKDD* 探索通讯，19(1):37–45，2017年 9 月。Minmin Chen、Alex Beutel、Paul Covington、Sagar Jain、Francois Belletti 和 Ed H. Chi。用于 REINFORCE 推荐系统的 Top-k 离策略校正。载于*WSDM*，第 456–464 页，2019年。Paolo Cremonesi 和 Dietmar Jannach。推荐系统研究的进展：危机？什么危机？《人工智能杂志》，42(3):43–54，2021年 11 月。

Romain Deffayet、Jean-Michel Renders 和 Maarten de Rijke。评估点击模型对策略分布偏移的鲁棒性。*ACM* 信息系统汇刊，2022年10月。

Maurizio Ferrari Dacrema、Paolo Cremonesi 和 Dietmar Jannach。我们真的取得了很大进展吗？对近期神经推荐方法的一项令人担忧的分析。载于 *RecSys*，第 101–109 页，2019年。Justin Fu、Aviral Kumar、Matthew Soh 和 Sergey Levine。诊断深度 Q 学习算法的瓶颈。载于 *ICML*，第 2021–2030 页，2019年。Justin Fu、Aviral Kumar、Ofir Nachum、George Tucker 和 Sergey Levine。D4RL：深度数据驱动强化学习数据集。*arXiv:2004.07219*，2020年。Justin Fu、Mohammad Norouzi、Ofir Nachum、George Tucker、Ziyu Wang、Alexander Novikov、Mengjiao Yang、Michael R. Zhang、Yutian Chen、Aviral Kumar、Cosmin Paduraru、Sergey Levine 和 Tom Le Paine。深度离策略评估基准测试。载于 *ICLR*，2021年。Scott Fujimoto、David Meger 和 Doina Precup。无需探索的离策略深度强化学习。载于 *ICML*，2019年。Florent Garcin、Boi Faltings、Olivier Donatsch、Ayar Alazzawi、Christophe Bruttin 和 Amr Huber。在 swissinfo.ch 对新闻推荐系统进行离线和在线评估。载于 *RecSys*，第 169–176 页，2014年。Dibya Ghosh、Anurag Ajay、Pulkit Agrawal 和 Sergey Levine。离线强化学习策略应经过训练以具备自适应能力。载于 *ICML*，第 7513–7530 页，2022年。

Carlos A. Gomez-Uribe 和 Neil Hunt。《Netflix 推荐系统：算法、商业价值与创新》。*ACM 信息系统管理汇刊*，6(4)，2016年12月。

GroupLens。MovieLens 数据集。网址 https://grouplens.org/datasets/movielens/。Pengjie Gu、Mengchen Zhao、Chen Chen、Dong Li、Jianye Hao 和 Bo An。为离线强化学习学习基于伪度量的动作表示。收录于 *ICML*，第 7902–7918 页，2022年。Caglar Gulcehre、Ziyu Wang、Alexander Novikov、Tom Le Paine、Sergio G´omez Colmenarejo、Konrad Zo-lna、Rishabh Agarwal、Josh Merel、Daniel Mankowitz、Cosmin Paduraru、Gabriel Dulac-Arnold、Jerry Li、Mohammad Norouzi、Matt Hoffman、Nicolas Heess 和 Nando de Freitas。RL Unplugged：离线强化学习基准测试套件。收录于 *NeurIPS*，2020年。Josiah Hanna、Scott Niekum 和 Peter Stone。使用估计的行为策略进行重要性采样策略评估。收录于 *ICML*，第 2605–2613 页，2019年。Bal´azs Hidasi、Alexandros Karatzoglou、Linas Baltrunas 和 Domonkos Tikk。基于循环神经网络的基于会话的推荐。收录于 *ICLR*，2016年。Jin Huang、Harrie Oosterhuis、Maarten de Rijke 和 Herke van Hoof。避免数据集偏差影响仿真：用于基于强化学习的推荐系统的去偏模拟器。收录于 *RecSys*，第 190–199 页，2020年。

Ahmed Hussein、Mohamed Medhat Gaber、Eyad Elyan 和 Chrisina Jayne。模仿学习：学习方法综述。*ACM Comput. Surv.*，50(2)，2017年4月。Instacart。Instacart 购物篮分析，2017年。网址 https://www.kaggle.com/c/instacart-market-basket-analysis/data。Dietmar Jannach、Paul Resnick、Alexander Tuzhilin 和 Markus Zanker。推荐系统——超越矩阵补全。*Commun. ACM*，59(11):94–102，2016年10月。Olivier Jeunen。重新审视隐式反馈推荐系统的离线评估。收录于*Rec-Sys*，第 596–600 页，2019年。Olivier Jeunen 和 Bart Goethals。用于推荐离策略学习的悲观奖励模型。收录于 *RecSys*，第 63–74 页，2021年。Luo Ji、Qi Qin、Bingqing Han 和 Hongxia Yang。通过强化学习优化冷启动推荐中的生命周期价值。收录于 *CIKM*，第 782–791 页，2021年。Yitong Ji、Aixin Sun、Jie Zhang 和 Chenliang Li。对推荐系统离线评估中数据泄漏问题的批判性研究。*arXiv:2010.11060*，2020年。Thorsten Joachims、Adith Swaminathan 和 Tobias Schnabel。利用有偏反馈进行无偏学习排序。收录于 *WSDM*，第 781–789 页，2017年。Haruka Kiyohara 和 Kosuke Kawakami。OFRL：从实践角度设计离线强化学习与策略评估平台。收录于 *CONSE-QUENCES+REVEAL@RecSys*，2022年。Ilya Kostrikov、Ashvin Nair 和 Sergey Levine。基于隐式Q学习的离线强化学习。*arXiv:2110.06169*，2021年。Walid Krichene 和 Steffen Rendle。关于物品推荐的采样指标。收录于 *KDD*，第 1748–1757 页，2020年。Aviral Kumar、Anikait Singh、Stephen Tian、Chelsea Finn 和 Sergey Levine。离线无模型机器人强化学习工作流。收录于 *CoRL*，2021年。Last.fm。网址 https://last.fm/api。Hojoon Lee、Dongyoon Hwang、Kyushik Min 和 Jaegul Choo。探索交互式推荐系统中长期用户反馈的验证。收录于*SIGIR*，第 2607–2611 页，2022年。Sergey Levine、Aviral Kumar、George Tucker 和 Justin Fu。离线强化学习：教程、综述及对开放问题的展望。*arXiv:2005.01643*，2020年。Malte Ludewig、Noemi Mauro、Sara Latifi 和 Dietmar Jannach。基于会话的推荐中神经方法与非神经方法的性能比较。收录于 *RecSys*，第 462–466 页，2019年。

Sean M. McNee、John Riedl 和 Joseph A. Konstan。准确还不够：准确率指标如何损害推荐系统。载于 *CHI*，第 1097–1101 页，2006年。Harrie Oosterhuis 和 Maarten de de Rijke。反事实学习排序中的鲁棒泛化与安全查询专门化。载于 *WWW*，第 158–170 页，2021年。OpenAI、Marcin Andrychowicz、Bowen Baker、Maciek Chociej、Rafal Jozefowicz、Bob McGrew、Jakub Pachocki、Arthur Petron、Matthias Plappert、Glenn Powell、Alex Ray、Jonas Schneider、Szymon Sidor、Josh Tobin、Peter Welinder、Lilian Weng 和 Wojciech Zaremba。灵巧手内操作学习。《国际机器人研究杂志》，39(1):3–20，2020年。Doina Precup、Richard S. Sutton 和 Satinder P. Singh。用于离策略策略评估的资格迹。载于 *ICML*，第 759–766 页，2000年。Rongjun Qin、Songyi Gao、Xingyuan Zhang、Zhen Xu、Shengkai Huang、Zewen Li、Weinan Zhang 和 Yang Yu。NeoRL：用于离线强化学习的近真实世界基准。 *arXiv:2102.00714*，2021年。Massimo Quadrana、Paolo Cremonesi 和 Dietmar Jannach。序列感知型推荐系统。 *ACM Comput. Surv.*，51(4)，2018年7月。Alfredo Reichlin、Giovanni Luca Marchetti、Hang Yin、Ali Ghadirzadeh 和 Danica Kragic。回归流形：从分布外状态中恢复。 *arXiv:2207.08673*，2022年。Steffen Rendle、Li Zhang 和 Yehuda Koren。评估基线方法之难：一项关于推荐系统的研究。 *arXiv:1905.01395*，2019年。RetailRocket。RetailRocket 推荐系统数据集，2016年。网址 https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset。Jing-Cheng Shi、Yang Yu、Qing Da、Shi-Yong Chen 和 Anxiang Zeng。Virtual-taobao：为强化学习虚拟化真实世界在线零售环境。载于 *AAAI*，第 4902–4909 页，2019年。Aixin Sun。从反直觉观察到重新审视推荐系统。*arXiv:2210.04149*，2022年。Zhu Sun、Di Yu、Hui Fang、Jie Yang、Xinghua Qu、Jie Zhang 和 Cong Geng。我们的评估是否严谨？面向可复现评估与公平比较的推荐基准测试。载于*RecSys*，第 23–32 页，2020年。Richard Sutton 和 Andrew Barto。《强化学习：导论》。麻省理工学院出版社，2018年。Adith Swaminathan 和 Thorsten Joachims。通过反事实风险最小化，从记录的老虎机反馈中进行批量学习。《机器学习研究杂志》，16(52):1731–1755，2015年。

Josh Tobin、Rachel Fong、Alex Ray、Jonas Schneider、Wojciech Zaremba 和 Pieter Abbeel。通过领域随机化，将深度神经网络从仿真迁移到真实世界。载于*2017*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），第 23–30 页，2017年。doi：10.1109/IROS.2017.8202133。

Hado van Hasselt、Yotam Doron、Florian Strub、Matteo Hessel、Nicolas Sonnerat 和 Joseph Modayil。深度强化学习与致命三元组。*arXiv:1812.02648*，2018年。

Wayne Xin Zhao、Zihan Lin、Zhichao Feng、Pengfei Wang 和 Ji-Rong Wen。重新审视 Top-N 推荐算法的适当离线评估。*ACM* 信息系统汇刊，2022年6月。