基于随机反馈建模长期用户参与度

Guoxi Zhang 京都大学信息学研究生院 日本京都
guoxi@ml.ist.i.kyoto-u.ac.jp

Xing Yao 中国证券登记结算有限责任公司 中国北京
854352859@qq.com

Xuanji Xiao*
虾皮科技 中国深圳
growj@126.com

arXiv:2302.06101v2 [cs.IR] 5 Apr 2023

摘要

推荐系统（RS）的终极目标之一是提升用户参与度。强化学习（RL）是实现这一目标的一种有前景的范式，因为它直接优化序列推荐的整体性能。然而，许多现有的基于强化学习的方法会带来巨大的计算开销，因为它们不仅需要存储推荐物品，还需要存储所有其他候选物品。本文提出了一种高效的替代方法，无需存储候选物品。其思路是直接从数据中建模用户参与度与物品之间的相关性。此外，所提方法考虑了用户反馈和终止行为中的随机性，这些随机性在推荐系统中普遍存在，但在以往基于强化学习的研究中鲜少讨论。通过在真实推荐系统上开展在线 A/B实验，我们验证了所提方法的有效性，以及对这两类随机性进行建模的重要性。

CCS 概念

• **计算方法** → **马尔可夫决策过程**;
• **信息系统** → 内容排序；个性化。

关键词

推荐系统、强化学习

ACM 参考格式：
Guoxi Zhang、Xing Yao 和 Xuanji Xiao。2023 年。基于随机反馈建模长期用户参与度。载于*2023*年*ACM Web*会议（*WWW '23 Companion*）伴随论文集，*2023*年*4*月*30*日至*5*月*4*日，美国得克萨斯州奥斯汀。 ACM，纽约州纽约，美国，5 页。https://doi.org/10.1145/3543873.3587626

1 引言

推荐系统（RS）是帮助用户发现有吸引力物品的软件系统。具体而言，序列推荐系统允许用户不断消费新物品，通常只需在推荐系统界面上向下滚动即可。序列推荐系统的主要目标是提升长期用户参与度。在这方面，短视的决策方式（例如仅根据预估点击率对物品进行排序）可能并非最优。

*通讯作者

允许免费出于个人或课堂用途制作本作品全部或部分内容的数字或纸质副本，但前提是不得为盈利或商业利益而制作或分发副本，且副本须保留本声明及首页上的完整引文。对于本作品中由作者以外的其他人拥有版权的部分，必须尊重其版权。注明出处后可以摘录。以其他方式复制、重新出版、发布到服务器或重新分发至列表，均须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。
WWW '23 Companion，*2023*年*4*月*30*日至*5*月*4*日，美国得克萨斯州奥斯汀© 2023 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-9419-2/23/04...$15.00 https://doi.org/10.1145/3543873.3587626

（pCTR）可能并非最优。例如，新闻的预估点击率可能高于情景喜剧花絮，但情景喜剧花絮能促使用户更多地观看有关该情景喜剧及其演员的视频。

强化学习（RL）已成为优化长期用户参与度的一项颇具前景的研究方向 [2, 4,12, 14]。借助强化学习，序列推荐被建模为智能体与环境之间的交互。在每一步中，智能体都会接收一个状态并选择一个动作。通常，状态是用户画像及其历史行为，而动作则是可供推荐的物品。每次选择动作后，智能体都会收到一个标量奖励。强化学习的目标是学习一个策略来选择动作，使智能体在与环境交互期间最大化所获得奖励之和。因此，如果将物品效用（例如点击）作为奖励，强化学习就是一种极具吸引力的范式，因为它能够自然地优化序列推荐系统的长期效用。

然而，为工业推荐系统训练强化学习模型可能颇具挑战。为了解这一问题，我们来看一个极简示例：智能体需要从一组候选物品中为用户选择一件物品。要训练预估点击率模型，需要记录用户、所选物品以及物品反馈的信息。然而，要训练诸如著名的深度Q网络（DQN） [6], 这样的强化学习模型，还需要额外记录候选物品集。由于候选集通常包含数百件物品，这会给工业应用带来巨大的开销。虽然可以通过使用物品嵌入向量来学习强化学习模型，从而降低开销，但据报道，这种方式的性能远不如直接使用物品来学习强化学习模型 [8]。

本文提出了一种高效的替代方案，用于优化长期用户参与度。所提方法的核心观察是，在工业场景中，推荐系统会定期使用由其他推荐系统生成的日志进行训练，后文将其称为行为策略。对于服务数百万用户的应用，行为策略通常经过了高度优化，因此，通过挖掘用户参与度与该策略所选物品之间的关系，我们可以获得一个用户参与度模型。值得注意的是，所提方法在学习过程中不需要候选物品集，因此可以轻松集成到任何现有的短视推荐系统中，而不会带来巨大的额外开销。

除了效率问题，所提方法还考虑了奖励和交互终止中的随机性。深度Q网络等强化学习算法通常假设：(a) 奖励由状态和动作确定；(b) 交互会持续无限步。这两项假设在推荐系统中都很少成立。现实中的推荐系统需要服务兴趣各异的用户，但对用户掌握的信息有限。考虑两个特征值相同的不同用户。由于他们的兴趣不同，他们对同一物品的反馈也会不同。从

推荐系统的角度来看，这意味着相同状态和动作对应的奖励往往具有随机性。出于同样的原因，他们与推荐系统交互的步数也不太可能相同。遗憾的是，迄今为止，这两类随机性都尚未得到解决。为对此前一种情况进行建模，本文采用了近期提出的分布式强化学习框架 [1]。针对后一种情况，本文将分布式强化学习框架扩展到随机终止场景，并学习一个模型来预测用户是否会终止交互。

我们在一个服务数百万用户的真实工业短视频推荐系统上评估了所提方法。在为期一周的在线 A/B 测试中，部署所提方法后，用户平均观看的视频数量提升了 2.72%，视频观看的平均时长提升了 1.59%。此外，实验还证实，对随机性进行建模对于有效建模长期用户参与度至关重要。

本文的贡献可总结如下。

- 本文提出了一种简单高效的方法，用于对推荐系统中的长期用户参与度进行建模。
- 本文提出，在对用户参与度进行建模时，应考虑用户反馈和终止的随机性。
- 本文在工业推荐系统上评估了所提方法，并证实了其有效性。

本文其余部分安排如下：第 2 节简要回顾相关文献，第 3 节介绍技术背景，第 4 节阐述所提方法，第 5 节展示在真实推荐系统上进行的实验，第 6 节对本文作总结。

2 相关工作

现有基于强化学习的推荐方法可分为基于策略梯度的方法 [2] 和深度Q网络算法 [13, 14] 的变体。本文旨在解决效率问题，可视为一种面向资源受限场景的补充方法。

在推荐系统领域，研究人员已在神经网络架构方面投入了大量精力。所提方法采用 DLRM [7]架构，但也可以与其他架构结合使用。同时，为了对奖励中的随机性进行建模，所提方法采用了近期提出的分布式强化学习框架 [1, 3]。至于随机终止，文献中已有研究在基于模型的强化学习 [9] 中考虑了这一问题，并采用进化算法 [11] 或元梯度下降 [10]进行建模。我们提出的方法与这些方法不同，它通过用户的离开行为来学习折扣因子。

3 预备知识

使用强化学习建模序列推荐。 用户与推荐智能体之间的序列交互被建模为一个折扣无限时域马尔可夫决策过程（MDP）：$<\mathcal{S}, \mathcal{A}, R, P, \gamma>$。状态 $\mathcal{S}$ 表示有关用户的信息，包括兴趣标签以及过去几分钟、几小时或几天内消费过的物品列表等特征。动作 $\mathcal{A}$ 是候选物品。在分布式强化学习中，奖励 R 被视为随机变量。本文将点击信号用作奖励，因此，若 $r = 1$ 实现为真，则表示相应物品被点击；若 $r = 0$ 则表示未被点击。转移动态 P 支配状态转移，而折扣因子 γ 用于定义价值分布。

动作价值分布。 我们关注对智能体与环境交互过程中获得的奖励总和进行建模，该总和刻画了长期用户参与度。由于我们假设奖励是随机变量，因此状态-动作价值 $Z^\pi(s, a)$——即在某个状态 s 下选择动作 a 并随后遵循 π 所获得的折扣奖励总和——也是随机变量。Bellemare 等人证明，$Z^\pi(s, a)$ 是分布式贝尔曼算子 $\mathcal{T}^\pi$ 的定点，该算子定义为 [1]：

$$\forall (s,a) \in \mathcal{S} \times \mathcal{A}, \mathcal{T}^\pi Z^\pi(s,a) \overset{D}{=} R(s,a) + \gamma Z^\pi(s', \pi(s')), \quad (1)$$

其中 $s' \sim P(s|s, a)$。$X \overset{D}{=} Y$ 表示随机变量 X 与随机变量 Y 同分布。

Dabney 等人提出使用所谓的分位数分布 [3]对 Z^π 进行参数化。其思路是学习一个参数化函数 $\beta: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^M$，以估计 Z^π 的 M 个均匀间隔分位数。给定来自 π 的数据，他们提出通过最小化分位数 Huber 损失来学习 Z^π：

$$L(\beta) = \mathbb{E}_{(s,a,r,s')\sim\pi} \frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{M} [\rho^\kappa_{\hat{\tau}_i}(r + \gamma\beta'(s', \pi(s'))_j - \beta(s,a)_i)], \quad (2)$$

其中，τ_m 是价值分布的第 i 个分位数，而 $\hat{\tau}_m = \frac{1}{2}(\tau_{m-1} + \tau_m)$（对于 $m = 1, 2, \ldots, M$）是分位数中点。为简单起见，τ_0 定义为 0，τ_M 定义为 1。$\rho^\kappa_\tau(u) = |\tau - \mathbb{I}(u < 0)| \mathcal{L}_\kappa(u)$，且 $\mathbb{I}(\cdot)$ 是指示函数。$\kappa > 0$ 是一个超参数，实践中我们使用 $\kappa = 1$。β' 与 β 具有相同的神经网络结构，其参数会定期从 β 复制。

4 所提方法

4.1 学习问题

我们假设智能体从由行为策略 b 生成的一些离线数据 D 中学习。D 是用户与智能体之间序列交互的日志。我们将用户相关信息表示为状态，将推荐物品表示为动作。如果相应物品被点击，则该决策的奖励为 1，否则为 0。最后，我们将 D 组织为转移，可写作 (s, a, r, s')，其中 s 和 s' 是两个连续的状态，a 是在 s 时选择的动作，r 是该决策的奖励。

我们的目标是学习一个函数 $g(s, a): \mathcal{S} \times \mathcal{A} \to \mathbb{R}$，用于预测在步骤 s 选择 a 并随后遵循 b 所获得的奖励总和，该总和刻画了选择 a 所带来的长期用户参与度。借助 $g(s, a)$，我们可以根据物品带来的长期用户参与度对其进行排序。

4.2 所提方法

所提方法基于以下观察。现实世界中的推荐系统通常同时采用基于学习的方法和基于规则的方法进行高度优化，并且会定期更新，以适应新的用户和新的物品。因此，用户往往会依次与其交互，从而使由此产生的数据包含物品对用户参与度影响的信息。因此，我们提出从数据中挖掘此类信息。

同时，正如第 1 节所述，用户反馈本质上具有随机性。因此，当将用户参与度建模为累积用户反馈时，我们必须适当地处理这种随机性。所提方法的核心思想是从数据 D 中估计行为策略 b 的 $Z^b(s, a)$ 价值分布。利用公式 2 可以实现这一点。借助 $Z^b(s, a)$，我们可以将其期望值用作长期用户参与度的评分函数，即 $g(s, a) = \mathbb{E}[Z^b(s, a)] = \frac{1}{M}\sum_j^M \beta(s, a)_j$。

现在我们讨论如何处理随机终止。在马尔可夫决策过程中，折扣因子 γ 用于定义价值函数和价值分布，但也可以将其解释为在状态-动作对 [5]之后交互继续进行的概率。因此，$1-\gamma$ 恰好是交互终止的概率。我们的想法是利用数据估计这一概率，并基于该估计学习 $Z^b(s, a)$。因此，$Z^b(s, a)$将能够考虑决策与交互终止之间的相关性。由此得到一个新的分布式贝尔曼算子 T_ℓ^π:

$$\forall (s,a) \in \mathcal{S}\times\mathcal{A}, \mathcal{T}_\ell^\pi Z^\pi(s,a) \overset{D}{=} R(s,a) + \gamma'(s', \pi(s'))Z^\pi(s', \pi(s')), \tag{3}$$

其中 $\gamma'(s, a) := \min(1-\ell(s, a), \eta)$。$\ell(s, a)$ 是交互在(s, a)之后终止的概率，而 $\eta \in (0, 1)$是一个超参数。在该算子中，下一状态 $Z^\pi(s', \pi(s'))$的价值分布由 $\gamma'(s, a)$加权，该项表示交互在$(s', \pi(s'))$之后继续进行的概率。因此，使用该算子学习到的 $Z^b(s, a)$将能够对随机终止进行建模。

关于 T_ℓ^π 的一个潜在问题是其收敛性。接下来，我们基于 [1] 对 T^π 的分析，证明在表格情形下 T_ℓ^π 是一个压缩映射。首先，设 F_X 和 F_Y 分别是随机变量 X 和 Y 的累积分布函数（CDF）。F_X 与 F_Y 之间的 Wasserstein 度量 d_p 可写为 $d_p(F_X, F_Y) = \mathrm{k}F_X(u) - F_Y(u)\mathrm{k}_p$[1]。对于两个动作价值分布 Z_1 和 Z_2，定义 $\bar{d}_p(Z_1, Z_2) := \sup_{s,a} d_p(Z_1(s, a), Z_2(s, a))$。我们的分析基于 d_p [1]的以下两个性质。假设随机变量 A 与 X 和 Y 相互独立，且 a 是常数，则:

$$\begin{aligned} d_p(A + X, A + Y) &\le d_p(X, Y), \\ d_p(aX, aY) &\le |a| d_p(X, Y). \end{aligned} \tag{4}$$

假设我们有来自策略 π 的样本。那么，对于 S × A 上的两个分布 Z_1 和 Z_2，以及任意状态 $s \in \mathrm{S}$ 和 $a \in \mathrm{A}$，应用 T_ℓ^π 后两者之间的 Wasserstein 度量可表示为:

$$\begin{aligned} & d_p(\mathcal{T}_\ell^\pi Z_1, \mathcal{T}_\ell^\pi Z_2) \\ &= d_p(R(s,a) + \gamma' Z_1(s', \pi(s')), R(s,a) + \gamma' Z_2(s', \pi(s')) \\ &\le d_p(\gamma'(s', \pi(s'))Z_1(s', \pi(s')), \gamma'(s', \pi(s'))Z_2(s', \pi(s'))) \\ &\le \eta d_p(Z_1(s', \pi(s')), Z_2(s', \pi(s'))) \\ &\le \eta \sup_{s', a'} d_p(Z_1(s', a')), Z_2(s', a')) \\ &= \eta \bar{d}_p(Z_1, Z_2) \end{aligned} \tag{5}$$

因此 $\bar{d}_p(\mathrm{T}_\ell^\pi Z_1, \mathrm{T}_\ell^\pi Z_2) = \sup_{s,a} d_p(\mathrm{T}_\ell^\pi Z_1, \mathrm{T}_\ell^\pi Z_2) \le \eta \bar{d}_p(Z_1, Z_2)$，这意味着 T_ℓ^π 是 $\bar{d}_p$ 中的 η-压缩映射。

[1]为简化记号，在可能的情况下，将随机变量及其累积分布函数视为同一对象。

算法 1： 所提方法的实用算法框架）

输入： T_c，将 β 复制到 β'的复制频率
B, 小批量的大小
随机初始化 β，并将其值复制到 β'
针对小批量 $i = 1, 2, \ldots, \frac{n}{B}$ **do**
  使用公式 6 更新 β 和 ℓ。
  如果 i 取模 $T_c = 0$ **那么**
    $\beta' = \beta$
  end
end

4.3 实用算法

对于实用算法，我们使用数据 D 同时估计 $Z^b(s, a)$ 和 $\ell(s, a)$。待最小化的目标函数如下。

$$\begin{aligned} & L_\ell(\beta, \ell) = \\ & \frac{1}{|\mathcal{D}|} \sum_{(s,a,r,s') \in \mathcal{D}} \{ \frac{1}{M} \sum_{i=1,j=1}^{M} [\rho_{\hat{\tau}_i}^\kappa (r + \gamma' \beta'(s', \pi(s'))_j - \beta(s,a)_i)] \\ & - e \log \ell(s,a) - (1-e)\log(1-\ell(s,a))\}. \end{aligned} \tag{6}$$

$e = 1$ 如果交互在（s, a）时终止，则.....；否则.....。β和 ℓ可以参数化为一个多目标网络，其中它们共享特征嵌入，但使用不同的前馈网络。算法 1 给出了该算法的伪代码。

5 实验

本节在一个工业推荐系统上评估所提方法，该系统会生成“你可能也喜欢的视频”信息流。具体而言，我们研究以下两个问题。

(1) 我们能否通过建模 Z^b（s, a）来提升真实世界推荐系统的用户参与度？(2) 在实践中，对奖励和终止中的随机性进行建模是否有益?

5.1 方法论

我们遵循 A/B 测试的标准流程。我们将用户随机哈希分配到对照组和测试组。在评估期开始前，我们通过统计检验确认对照组和测试组之间不存在显著差异。每个组至少有 630,000 名用户。

由于我们的目标是通过对长期用户参与度进行建模来改进推荐系统，因此我们在将算法部署到系统后评估了推荐系统的性能。令 $g^b(s, a)$ 为推荐系统当前的排序函数。那么，在部署用户参与度估计器 $g(s, a)$ 后，新的排序函数变为 $g^b\mathrm{new}(s, a) = g^b(s, a) + w * g(s, a)$，其中 w 是在初步实验中调优的参数。

我们使用视频消费指标作为用户参与度的评估指标，具体列于表 1。我们报告将算法部署到推荐系统后性能的变化。正向变化表明，某种算法能够提升当前在线策略的性能。

表 1：在线实验中使用的评估指标。VV 是用户点击的视频数量，IMP 是向用户展示的视频数量。DUR 是用户观看视频所花时间的总和。这三项指标衡量 A/B 测试期间的整体表现，因此我们将它们视为用户参与度指标。点击率和转化率是单个视频的指标。它们反映用户对单个物品的喜好程度，因此是短期效用指标。

指标	含义
视频观看次数（VV）	#点击次数 / #用户
#曝光次数 (IMP)	#曝光次数 / #用户
时长 (DUR)	Í(播放时长) / #用户
CTR	#点击次数 / #曝光次数
CVR	Í(播放时长) / Í(点击视频时长)

5.2 算法

对照组。 一种高度优化的短视频推荐系统，应用于某些工业应用中。

所提方法。 所提方法在建模用户参与度时，考虑了奖励和终止的随机性。

随机奖励。 所提方法的一种变体考虑了奖励的随机性，但不考虑终止的随机性。将此方法称为 RR。

完全确定性。 所提方法的一种变体直接估计 $g(s,\ a)$，而不考虑奖励和终止的随机性。将此方法称为 FD。

5.3 实现细节

实验中使用了 DLRM 结构。总共有 40 个特征，包括用户 ID、物品 ID 和物品关键词。这些离散特征被映射为 $\mathbb{R}^{32}$ 中的嵌入向量。每经过 100 个梯度步数，目标模块的参数就会从策略模块复制一次。算法使用单个 P40 GPU 每天进行训练。我们将 Adam 用作这三种算法的优化器，学习率设为 0。00015。在对奖励的随机性进行建模时，分位数的数量设为 200。

5.4 结果

表 2 展示了在线评估的结果。FD 未能取得优于对照组的显著提升。RR 显著提升了 VV、点击率和转化率，但其对 DUR 的提升并不显著。这些结果表明，该组用户可能比对照组观看了略多的视频。此外，结果还表明，对奖励中的随机性进行建模确实有助于刻画物品对用户参与度的影响。与此同时，所提方法显著提升了 VV、DUR、点击率和转化率，表明终止过程中的随机性也是对用户参与度建模的另一个关键因素。总的来说，表 2 证实，所提方法可作为一种高效的替代方案，用于对推荐系统中的长期用户参与度进行建模。

表 2：在线 A/B 测试结果。括号中的数字为 p 值。在未考虑任何随机性的情况下，FD 未能显著提升任何指标。考虑奖励中的随机性后，RR 显著提升了 VV、点击率和转化率。所提方法显著提升了 VV、DUR、点击率和转化率。结果证实，对奖励和终止过程中的随机性进行建模对于用户参与度建模至关重要。

指标	FD	RR	所提方法
VV	1.19% (0.11)	**1.58% (0.031)**	**2.72% (**2.1×10^{-4}**)**
IMP	0.140% (0.86)	-0.265% (0.74)	0.502% (0.53)
DUR	0.263% (0.73)	1.24% (0.10)	**1.59% (**3.5×10^{-2}**)**
CTR	1.05% (0.93)	**1.85% (**3.6×10^{-3}**)**	**2.21% (**4.1×10^{-4}**)**
CVR	0.410% (0.14)	**1.22% (**3.0×10^{-6}**)** </style)	**1.07% (**4.8×10^{-5}**)**

6 结论

本文探讨了如何高效地为推荐系统建模长期用户参与度。所提方法基于这样一个观察：在工业应用中，行为策略（即生产环境中的推荐系统）通常经过了高度优化。因此，所提模型不再从头学习最优策略，而是尝试从离线数据中捕捉用户参与度与推荐项目之间的相关性。尽管无法保证达到最优，该方法具有较高的计算效率，并已证实对实际系统有效。此外，本文提出对奖励和终止中的随机性进行建模，并验证了这两个因素对于用户参与度建模的重要性。

参考文献

[1] Marc G. Bellemare、Will Dabney 和 Rémi Munos。2017年。强化学习的分布视角。载于第三十四届国际机器学习会议论文集。PMLR，澳大利亚悉尼，449–458。[2] Minmin Chen、Alex Beutel、Paul Covington、Sagar Jain、Francois Belletti 和 Ed H. Chi。2019年。强化推荐系统的 top-k 离策略校正。载于第十二届美国计算机协会国际网络搜索与数据挖掘会议论文集。美国计算机协会，澳大利亚维多利亚州墨尔本，456–464。[3] Will Dabney、Mark Rowland、Marc G. Bellemare 和 Rémi Munos。2018年。基于分位数回归的分布式强化学习。载于第三十二届 *AAAI* 人工智能会议论文集。美国路易斯安那州新奥尔良，2892–2901。[4] Eugene Ie、Vihan Jain、Jing Wang、Sanmit Narvekar、Ritesh Agarwal、Rui Wu、Heng-Tze Cheng、Tushar Chandra 和 CraigBoutilier。2019年。SlateQ：面向推荐集合的强化学习可处理分解。载于第二十八届国际人工智能联合会议论文集。国际人工智能联合会议组织，2592–2599。[5] Michael L. Littman。1994年。作为多智能体强化学习框架的马尔可夫博弈。载于第十一届国际机器学习会议。美国新泽西州新不伦瑞克，157–163。[6] Volodymyr Mnih 等。2015年。通过深度强化学习实现人类水平控制。《自然》，518，7540，529–533。[7] Maxim Naumov 等。2019年。用于个性化和推荐系统的深度学习推荐模型。（2019年）。arXiv：1906.00091 [计算机科学·信息检索]。[8] Kai Wang、Zhene Zou、Yue Shang、Qilin Deng、Minghao Zhao、Runze Wu、Xudong Shen、Tangjie Lyu 和 Changjie Fan。2021年。RL4RS：基于强化学习的推荐系统真实世界基准。（2021年）。arXiv：2110.11073 [计算机科学·信息检索]。[9] Martha White。2017年。统一强化学习任务规范。载于第三十四届国际机器学习会议论文集。PMLR，澳大利亚悉尼，3742–3750。

[10] Zhongwen Xu、Hado P van Hasselt 和 David Silver。2018年。元梯度强化学习。载于神经信息处理系统进展。Curran Associates, Inc. [11] Naoto Yoshida、Eiji Uchibe 和 Kenji Doya。2013年。具有状态依赖折扣因子的强化学习。载于*2013 IEEE* 第三届国际发展与学习及表观遗传机器人联合会议。IEEE，1–6。[12] Xiangyu Zhao、Changsheng Gu、Haoshenglun Zhang、Xiaobing Liu、Xiwang Yang 和 Jiliang Tang。2019年。推荐系统中的在线广告深度强化学习。*arXiv* 预印本 *arXiv:1909.03602*。

[13] Xiangyu Zhao、Long Xia、Liang Zhang、Zhuoye Ding、Dawei Yin 和 Jiliang Tang。2018 年。逐页推荐的深度强化学习。载于第十二届 *ACM* 推荐系统会议论文集。美国计算机协会，加拿大不列颠哥伦比亚省温哥华，95–103。[14] Guanjie Zheng、Fuzheng Zhang、Zihan Zheng、Yang Xiang、Nicholas Jing Yuan、Xing Xie 和 Zhenhui Li。2018年。Drn：用于新闻推荐的深度强化学习框架。载于*2018*年万维网会议论文集。国际万维网会议指导委员会，法国里昂，167–176。