E-Commerce Letters 电子商务评论, 2024, 13(4), 4885-4892
Published Online November 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.1341716

Hans汉斯

基于强化学习的智能推荐方法在电子商务中的应用研究

文　艺

贵州大学大数据与信息工程学院，贵州　贵阳

收稿日期：2024年8月26日；录用日期：2024年9月9日；发布日期：2024年11月26日

摘　要

推荐系统可以通过推荐用户的个性化项目来减轻信息过载问题。在诸如电子商务之类的现实世界推荐中，系统与其用户之间的典型交互是向用户推荐一页项目并提供反馈；然后系统推荐一页新的项目。为了有效地捕捉这种交互并进行推荐，我们需要解决两个关键问题：(1) 如何根据用户的实时反馈更新推荐策略；(2) 如何生成具有适当显示的项目页面，这对传统的推荐系统提出了巨大的挑战。本文研究了页面推荐问题，旨在同时解决上述两个问题。提出了一种基于深度强化学习的页面推荐框架，该框架能够根据用户的实时反馈，优化页面中的项目，使其在二维页面中得到适当的显示。实验证明了本文方法的有效性。

关键词

推荐系统，商品展示策略，顺序偏好，深度强化学习，Actor-Critic

Research on the Application of Intelligent Recommendation Method Based on Reinforcement Learning in E-Commerce

Yi Wen

College of Big Data and Information Engineering, Guizhou University, Guiyang Guizhou

Received: Aug. 26th, 2024; accepted: Sep. 9th, 2024; published: Nov. 26th, 2024

Abstract

Recommender systems can alleviate the issue of information overload by recommending personalized

文章引用: 文艺. 基于强化学习的智能推荐方法在电子商务中的应用研究[J]. 电子商务评论, 2024, 13(4): 4885-4892.
DOI: 10.12677/ecl.2024.1341716

items to users. In real-world recommendations, such as in e-commerce, the typical interaction between the system and its users involves recommending a page of items to the user and receiving feedback; the system then recommends a new page of items. To effectively capture this interaction and make recommendations, two key issues must be addressed: (1) how to update the recommendation strategy based on the user's real-time feedback; (2) how to generate item pages with appropriate display, which poses significant challenges to traditional recommender systems. This paper investigates the problem of page recommendation, aiming to address both issues simultaneously. A page recommendation framework based on deep reinforcement learning is proposed. It can optimize the items on the page according to the user's real-time feedback, ensuring they are appropriately displayed in a two-dimensional page layout. Experimental results demonstrate the effectiveness of the proposed method.

Keywords

Recommender System, Item Display Strategy, Sequential Preference, Deep Reinforcement Learning, Actor-Critic

1. 引言

在电子商务应用场景中，推荐系统已经成为不可或缺的工具[1]。当用户面对海量商品时，推荐算法通过分析用户的浏览历史、购买行为及其他偏好数据，精准地推荐最符合其需求的商品或服务，从而帮助用户在繁杂的选择中快速找到心仪的产品。这不仅提高了用户的购物体验，还显著增加了平台的转化率和用户粘性[2]。在实际应用中，推荐系统往往需要与用户进行多轮交互，当前的推荐结果会直接影响用户的购买决策，并对后续的推荐质量产生影响。例如，用户可能不会接受初始推荐的商品，但这些反馈信息能够帮助系统在后续的推荐中做出更为精准的建议[3]。这个过程揭示了电商推荐系统面临的两个核心挑战：1) 如何高效捕捉和更新用户的动态偏好，并根据实时反馈调整推荐策略；2) 如何根据用户偏好生成个性化的商品展示页面。

提出了几组方法来解决在线个性化商品推荐问题，包括基于内容的方法[4] [5]，基于协同过滤的方法[6]-[8]和混合方法[9]-[11]。最近，深度学习模型[12]-[14]作为对传统方法的扩展和集成，因其能够处理复杂的用户－项目互动而受到关注。然而，这些方法不能有效地解决我们在商品推荐问题中提出的两个核心挑战。

为了解决这个问题，一个潜在的解决方案是利用强化学习技术将多步建议建模为顺序决策问题。已经证明，强化学习有能力为许多复杂的问题做出最优的多步决策，例如玩 Atari 和围棋[12]。通过使用强化学习，可以学习智能代理在每个步骤中推荐合适的项目，从而最大化所有步骤的全局最优推荐性能[15]。强化学习(RL)来自动学习最佳推荐策略[16]。基于强化学习的推荐系统有两个主要优点。首先，他们能够在交互过程中根据用户的实时反馈不断更新策略，直到系统收敛到最佳策略，生成最适合用户动态偏好的推荐。其次，最优策略是通过最大化用户的预期长期累积回报来制定的；而大多数传统推荐系统的设计都是为了最大化推荐的即时回报[17]。

与现有的工作不同我们提出了一种页面式推荐系统，该系统不仅能够生成一组不同且互补的推荐项目，还能优化这些项目在二维页面上的显示策略，从而实现最大化的用户回报。这种方法克服了传统强

化学习方法在推荐相似项目时的局限性，特别是在处理二维页面布局和用户浏览行为时，提供了更为精准和有效的推荐。具体来说，我们提出了一种页面式推荐系统，它能够同时生成一组不同且互补的推荐项目，并设计出在二维页面上展示这些项目的最佳策略，以实现最大化的用户回报。传统的强化学习方法(如DQN)虽然能推荐Q值最高的项目，但往往导致推荐内容过于相似[18]。在实际应用中，推荐补充性物品往往比仅推荐类似物品能带来更高的用户满意度。例如，在新闻提要中，用户更倾向于浏览多种不同主题的内容。此外，页面布局在电子商务推荐中尤为重要，因为推荐页面通常是二维网格而非线性列表。研究表明，用户的浏览习惯也非线性，他们更倾向于按块扫描页面，而非从上到下逐行浏览[19]。因此，一个有效的页面式推荐系统不仅要生成互补的推荐项目，还需同步设计适应用户浏览习惯的展示策略。

综上所述，我们提出了一种基于强化学习的页面式推荐系统，该系统不仅能够生成一组多样且互补的推荐项目，还能优化它们在二维页面上的显示策略，从而最大化用户的整体回报。相比传统的推荐算法，我们的方法更好地应对了电子商务场景中多轮交互和复杂页面布局的挑战。通过整合用户的实时反馈，我们的系统能够动态调整推荐策略，以精准捕捉用户的偏好变化。此外，我们的方法还考虑了用户的浏览行为，通过在二维页面中合理排列推荐项目，提升了用户体验。这种创新性的方法为电子商务推荐系统的未来发展提供了新的思路，能够更有效地满足用户的需求，提升平台的整体效益。

2. 基于 Actor-Critic 方法介绍

2.1. 马尔可夫决策过程

我们将推荐任务建模为马尔可夫决策过程(MDP)，并利用强化学习(RL)自动学习最优推荐策略，该策略可以在交互过程中不断更新推荐策略，并通过最大化用户的预期长期累积奖励来制定最优策略。我们首先定义一个五元组$(\mathcal{S},\mathcal{A},\mathcal{P},\mathcal{R},\gamma)$。其中各个符号的定义如下：

1) 状态空间$\mathcal{S}$：状态$s\in\mathcal{S}$被定义为基于用户的浏览历史生成的用户的当前偏好，即用户浏览的项目及其相应的反馈；

2) 动作空间$\mathcal{A}$：动作$a=\{a^1,\cdots,a^M\}\in\mathcal{A}$是基于当前状态$s$向用户推荐一个包含$M$个商品的页面；

3) 奖励$\mathcal{R}$：推荐代理(recommender agent, RA)在状态s采取动作a之后，即在向用户推荐商品页面时，用户浏览这些项目并提供她的反馈。她可以跳过(不点击)、点击或购买这些物品，代理根据用户的反馈获得即时奖励$r(s,a)$；

4) 转换P：转换$p(s'|s,a)$定义了RA采取行动a时从s到s的状态转换；

5) 折现因子γ：$\gamma\in[0,1]$定义了我们衡量未来回报现值时的折现因子。特别是，当$\gamma=0$时，RA只考虑即时奖励。换句话说，当$\gamma=1$时，所有未来的奖励都可以完全计入当前行动的奖励中。

具体来说，我们的推荐任务建模为马尔可夫过程，其中推荐代理(recommender agent, RA)与环境E(或用户)在一系列的时间步长进行交互。在每个时间步，RA根据E的状态$s\in\mathcal{S}$采取动作$a\in\mathcal{A}$，并接收奖励$r(s,a)$ (即RA根据用户的当前偏好推荐一页物品，并接收用户的反馈)。作为动作a的结果，环境E利用转换$p(s'|s,a)$将其状态更新为s'。强化学习的目标是找到一个推荐策略$\pi:\mathcal{S}\to\mathcal{A}$，使推荐系统的累积回报最大化。

在推荐系统中应用深度强化学习时，面临动态且庞大的动作空间以及计算成本高的问题。传统的深度Q学习架构无法有效应对这些挑战。因此，本文采用了基于Actor-Critic框架的推荐策略，该架构能够在处理大规模和动态动作空间的同时，减少冗余计算，提供更优的推荐效果。Actor-Critic架构适合于大的动态动作空间，并且还可以同时减少冗余计算。

在Actor-Critic框架中，Actor架构输入当前状态s并旨在输出确定性动作(或推荐M个商品的页面)，

即 $s \to a = \left\{a^1, \cdots, a^M\right\}$。Critic 只输入当前状态－动作对，而不是所有潜在的状态－动作对，这避免了前面提到的计算成本，输出相应的 Q 值，并利用最优 Q 值找到最佳动作，Q 值函数计算如下所示：

$$Q(s,a) = \mathbb{E}_{s'}\left[r + \gamma \max_{a'} Q^*\left(s', a'\right) \middle| s, a\right] \tag{1}$$

其中 $Q^*\left(s', a'\right)$ 为预测下一时刻，状态 s' 条件下采取动作 a' 的最优 Q 值函数。Q 值函数是对所选择的动作是否与当前状态匹配的判断，即，推荐是否与用户的偏好匹配。最后，Actor 根据 Critic 的判断，按照提高推荐性能的方向更新自身的参数，以便在下一次迭代中输出更合适的行为。接下来，我们将详细介绍使用者和评论者体系结构。

2.2. Actor 框架的架构

在设计基于强化学习的推荐系统时，我们面临三个主要挑战：初始化偏好设置、实时学习偏好以及联合生成推荐和页面展示。针对这些挑战，我们提出了一个基于 Actor-Critic 架构的解决方案，该架构在推荐精度和用户体验之间取得了良好的平衡。

在每个推荐会话开始时，系统需要快速、准确地初始化用户的偏好，即使缺乏丰富的用户历史数据。为此，我们在架构中引入了 Encoder 部分，它采用带有门控循环单元(GRU)的循环神经网络(RNN)来编码用户的初始偏好信息。Encoder 通过处理用户的历史行为数据，生成偏好向量 h_t，该向量表示用户在时间步 t 时的偏好状态：

$$h_t = GRU\left(x_t, h_{t-1}\right) \tag{2}$$

其中，x_t 表示当前时间步的输入，h_{t-1} 表示前一个时间步的隐藏状态。通过这个过程，系统能够在初始阶段迅速捕捉用户的偏好，即使在数据稀缺的情况下，也能为接下来的推荐生成高相关性的内容。

在用户与系统的持续交互过程中，用户的偏好会不断变化。因此，系统必须具备实时学习和更新用户偏好的能力，以确保推荐内容的连续性和准确性。为了实现这一点，框架中的 Actor 部分利用从 Encoder 生成的偏好向量 h_t 来生成推荐项 y_t。该过程可表示为：

$$y_t = Actor\left(h_t, c_t\right) \tag{3}$$

其中，c_t 为上下文向量，包含当前会话中的用户交互信息。Actor 不仅生成推荐项，还根据实时反馈动态调整推荐内容，以适应用户的变化需求。此外，Critic 部分通过计算 Q 值来评估推荐策略的效果，确保系统在每个步骤中都能够基于最新的用户偏好提供最优推荐。Critic 通过以下 Bellman 方程对推荐策略进行优化：

$$Q\left(s_t, y_t\right) = r_t + \gamma \max Q\left(s_{t+1}, y_{t+1}\right) \tag{4}$$

其中，r_t 是即时回报，γ 是折扣因子，表示未来奖励的重要性。通过 Critic 的反馈，Actor 不断调整推荐策略，直至策略收敛到最优解。

通过使用 Actor-Critic 架构，系统能够处理动态和高维的动作空间。在这一过程中，推荐项 y_t 是通过连续嵌入来表示的，避免了传统方法中使用离散索引导致的局限。动作空间 A_t 随时间动态变化，因此系统在每一轮推荐中都需考虑新加入或已移除的项目。Critic 部分计算所有状态－动作对的 Q 值，并反馈给 Actor，用于调整布局策略，从而实现最优的页面展示效果。最终，结合这些策略，Actor 框架能够生成一组不同且互补的推荐项目，并同时优化其在二维页面上的展示策略，从而显著提升用户体验。

2.3. Critic 框架的架构

在 Actor-Critic 框架中，Critic 部分的作用是估计当前策略的 Q 值，并通过最小化 Q 值误差来指导

Actor 选择最优动作。Critic 网络的输入包括状态和动作对(s,a)，输出为对应的 Q 值$Q(s,a)$。Critic 网络的更新基于 Bellman 方程，具体公式如公式(1)所示。Critic 通过最小化以下损失函数来更新网络参数：

$$L(\theta)=\mathbb{E}_{(s,a,r,s')}\left[\left(Q_\theta(s,a)-y\right)^2\right] \tag{5}$$

其中，$y=r+\gamma\max_{a'}Q_{\theta'}(s',a')$，$\theta'$是目标网络的参数。

综上述，Critic 的更新过程为：1、状态–动作对(s,a)的输入：Critic 网络接收当前状态 s 和动作 a，并输出估计的 Q 值$Q(s,a)$。2、Q 值的计算：使用 Bellman 方程估计未来的期望回报，更新 Q 值。3、参数更新：通过最小化损失函数$L(\theta)$，逐步调整 Critic 网络的参数θ，使 Q 值估计更加准确。

在处理推荐系统中大规模动态动作空间的任务时，传统的深度 Q 学习方法通常难以有效应对这些挑战，因为这些方法在处理较小的固定动作空间(如 Atari 游戏)时表现较好，但在面对推荐系统中不断变化的动作空间时效果有限。而 Actor-Critic 架构通过将 Actor 与 Critic 结合，可以更好地适应动态环境，从而实现更优的推荐策略。

3. 算法设计

3.1. 训练过程描述

如 2.2 节所述，我们将用户的偏好s^{cur}映射到新的推荐页面(a^{cur})。在 M 个项目的页面中，a^{cur}包含 M 个项目的项目嵌入，即$\{e_1,\cdots,e_M\}$。然而，a^{cur}是一个原型动作，因为生成的项嵌入$e_i\in a^{cur}$可能不在现有的项嵌入空间 I 中。因此，我们需要从原动作a_{pro}^{cur}映射到有效动作a_{val}^{cur}，其中$\left\{e_i\in I\middle|\forall e_i\in a_{val}^{cur}\right\}$。

当我们在在线环境中训练所提出的框架时，RA 可以通过在一系列时间步长上顺序地选择推荐项来与用户交互。因此，在在线环境中，RA 能够从用户接收针对推荐项目的实时反馈。在本文中，对于a_{pro}^{cur}中的每个e_i，我们选择最相似的$e_i\in I$作为a_{val}^{cur}中的有效项嵌入。在这项工作中，我们选择余弦相似度为：

$$e_i=\arg\max_{e\in I}\frac{e_i^{\mathrm{T}}\cdot e}{\|e_i\|\|e\|}=\arg\max_{e\in I}e_i^{\mathrm{T}}\cdot\frac{e}{\|e\|} \tag{6}$$

为了降低计算成本，我们预先计算$e/\|e\|$，并采用项目召回机制，以减少相关项目的数量。

我们提出了映射算法。Actor 首先生成原动作a_{pro}^{cur}。对于a_{pro}^{cur}中的每个e_m，RA 根据余弦相似性选择最相似的项目，然后将该项目添加到a_{val}^{cur}中与a_{pro}^{cur}中的e_m相同的位置。最后，RA 从项目嵌入空间中删除该项目，这可以防止在一个推荐页面中重复推荐相同的项目。然后，RA 向用户推荐具有瓦尔的新推荐页面，并接收来自用户的即时反馈(奖励)。奖励 r 是本页面所有项目的奖励总和：

$$r=\sum_{m=1}^{M}reward(e_m) \tag{7}$$

3.2. Deep-Recommendation 算法设计

在这项工作中，我们利用 DDPG 算法来训练所提出的 Actor-Critic 框架的参数。可以通过最小化损失函数$L(\theta^\mu)$的序列来训练批评者，如下所示：

$$L(\theta^\mu)=\mathbb{E}_{s,a,r,s'}\left[\left(r+\gamma Q_{\theta^{\mu'}}(s',f_{\theta^{\pi'}}(s'))-Q_{\theta^\mu}(s,a)\right)^2\right] \tag{8}$$

其中θ^μ表示 Critic 中的所有参数。Critic 是从存储在重放缓冲区中的样本中训练出来的。存储在重放缓冲器中的动作由有效动作a_{val}^{cur}生成，即$a=conv2d(a_{val}^{cur})$。这允许学习算法利用实际执行的动作的信息来训练 Critic。

等式(8)中的第一项 $y = r + \gamma Q_{\theta^{\mu'}}\left(s', f_{\theta^{\pi'}}\left(s'\right)\right)$ 是当前迭代的目标。在优化损失函数 $L\left(\theta^{\mu}\right)$ 时，来自前一次迭代的参数 $\theta^{\mu'}$ 是固定的。在实验中，通过随机梯度下降来优化损失函数通常在计算上是有效的，而不是计算上述梯度中的全部期望。损失函数 $L\left(\theta^{\mu}\right)$ 对参数 θ^{μ} 的导数如下所示：

$$\nabla_{\theta^{\mu}} L\left(\theta^{\mu}\right) = \mathbb{E}_{s,a,r,s'}\left[\left(r + \gamma Q_{\theta^{\mu'}}\left(s', f_{\theta^{\pi'}}\left(s'\right)\right) - Q_{\theta^{\mu}}\left(s,a\right)\right)\nabla_{\theta^{\mu}} Q_{\theta^{\mu}}\left(s,a\right)\right] \tag{9}$$

我们使用策略梯度更新 Actor：

$$\nabla_{\theta^{\mu}} f_{\theta^{\mu}} \approx \mathbb{E}\left[\nabla_{\hat{a}} Q_{\theta^{\mu}}\left(s,\hat{a}\right)\nabla_{\theta^{\mu}} f_{\theta^{\pi}}\left(s\right)\right] \tag{10}$$

其中 $\hat{a} = f_{\theta^{\pi}}\left(s\right)$，$\hat{a}$ 由原动作 a_{pro}^{cur} $\left(a = conv2d\left(a_{pro}^{cur}\right)\right)$生成。原动作 a_{pro}^{cur} 是 Actor 输出的实际动作。这保证了在策略 $f_{\theta^{\pi}}$ 的实际输出处采取策略梯度。

综上所述，我们提出了深度推荐(Deep-recommendation, DR)的在线训练算法。在每次迭代中，有两个阶段，即，1) 转换生成阶段，以及 2) 参数更新阶段。对于转换生成阶段：给定当前状态 s_t，RA 首先根据算法 1 推荐一页项 a_t；然后 RA 观察奖励 r_t 并将状态更新为 s_{t+1}；最后 RA 将转换 $\left(s_t, a_t, r_t, s_{t+1}\right)$ 存储到重放缓冲器 D 中。对于参数更新阶段：RA 对转换的小批量采样 $\left(s,a,r,s'\right)$，然后按照标准 DDPG 程序更新 Actor 和 Critic 的参数。

4. 实验分析

在本节中，我们进行了广泛的实验，从一个真实的电子商务公司的数据集，以评估所提出的框架的有效性。我们主要关注两个问题：(1) 与代表性基线相比，所提出的框架表现如何；以及(2) Actor 和 Critic 中的组件如何对性能做出贡献。

4.1. 实验设置

我们在 2023 年 9 月来自某电商公司的数据集上评估了我们的方法。我们随机收集了 1,000,000 个推荐会话(9,136,976 项物品)，按照时间顺序排列，并将前 70%的会话用作训练/验证集，后 30%的会话用作测试集。对于一个新的会话，初始状态来自用户的之前会话。在本研究中，我们利用 N = 10 个之前点击/购买的物品来生成初始状态。每次 RA 向用户推荐一页 M = 10 个物品(5 行 2 列)。跳过/点击/购买一个物品的奖励 r 分别设为 0、1 和 5。物品嵌入/类别嵌入/反馈嵌入的维度分别为|E| = 50、|C| = 35 和|F| = 15。我们将折扣因子 γ 设为 0.95，目标网络的软更新率 τ 设为 0.01。在线测试中，我们利用一个推荐会话中所有奖励的总和作为指标。

4.2. 性能分析

根据[20]，我们为在线测试构建了一个模拟在线环境(适应于我们的情况)。我们将 DR 与 DQN 和 DDPG 进行比较。在这里，我们采用在线训练策略来训练 DDPG 和 DR (均为 Actor-Critic 框架)。基准模型也可以通过模拟在线环境生成的奖励进行训练。我们使用不同于训练集的数据来构建模拟在线环境，以避免过拟合。

由于测试阶段基于模拟器，我们可以人工控制推荐会话的长度，以研究短会话和长会话的表现。我们定义短会话为包含 10 页推荐内容，长会话为包含 50 页推荐内容。结果如图 1 所示。我们注意到，DDPG 收敛速度的表现与 DQN 相似，但 DDPG 的平均奖励大于 DQN，本文提出的 DR 算法无论在收敛性还是平均奖励都优于 DDPG 和 DQN。这表明，Actor-Critic 框架适合具有巨大动作空间的实际推荐系统。

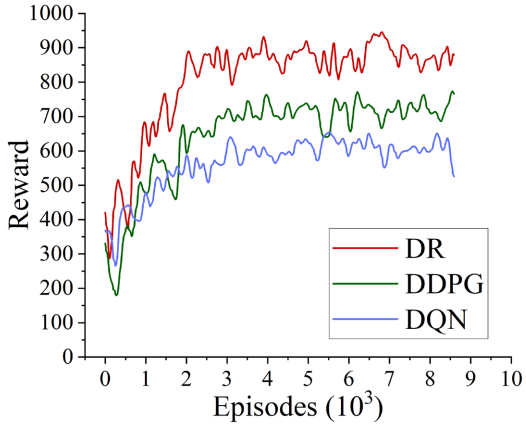

Figure 1. Convergence speed comparison of training of different algorithms
图 1. 不同算法的训练收敛速度对比

在图 2(a)短推荐会话中，DR 的表现优于 DQN 和 DDPG。在图 2(b)长推荐会话中，DR 的表现同样远优于传统的 DQN 和 DDPG。DR 能够学习用户的实时偏好，并优化推荐页面中的物品。

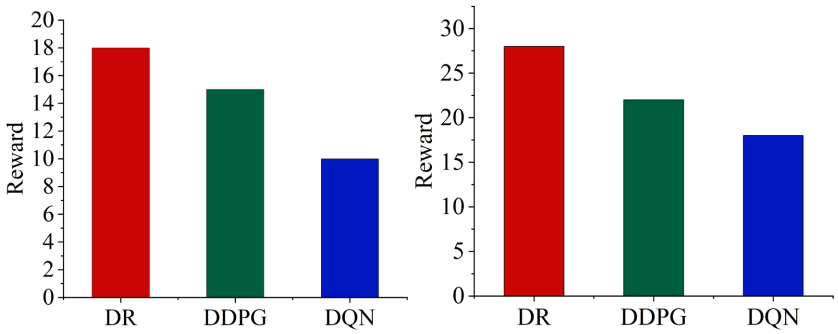

Figure 2. Comparison of recommendation performance of different algorithms
图 2. 不同算法的推荐性能对比

5. 总结

在本文中，我们提出了一种新的页面推荐框架 DR，它利用深度强化学习来自动学习最佳推荐策略，并同时优化一个页面的项目。我们验证了我们的框架的有效性与广泛的实验数据的基础上，从一个真实的电子商务公司。进一步的研究方向包括降低从原型动作到有效动作的映射的时间复杂度，并在一个强化学习框架中协同处理多个任务，如搜索，投标，广告和推荐。

参考文献

[1] 周艳榕. 基于个性化特征的电子商务智能推荐系统[J]. 现代电子技术, 2020, 43(19): 155-158, 162.

[2] 王利娥, 许元馨, 李先贤, 等. 移动商务推荐系统中的一种基于 P2P 的隐私保护策略[J]. 计算机科学, 2017, 44(9): 178-183.

[3] 李征, 黄雪原, 袁科. 基于位置社交网络的兴趣点推荐研究[J]. 计算机应用研究, 2022, 39(11): 3211-3219.

[4] 喻继军, 熊明华. 电子商务推荐系统公平性研究进展[J]. 现代信息科技, 2023, 7(14): 115-124.

[5] 李昌盛, 伍之昂, 张璐, 等. 关联规则推荐的高效分布式计算框架[J]. 计算机学报, 2019, 42(6): 1218-1231.

[6] 刘旭东, 陈德人, 钟苏丽. 使用群体兴趣偏好度的协同过滤推荐[J]. 计算机工程与应用, 2010, 46(34): 129-131.

[7] 王渊. 基于 RFM 模型的协同过滤方法及其在个性化推荐中的应用[D]: [硕士学位论文]. 杭州: 杭州电子科技大学, 2014.

[8] 李大学, 谢名亮, 赵学斌. 结合项目类别信息的协同过滤推荐算法[J]. 重庆邮电大学学报(自然科学版), 2010, 22(6): 823-827.

[9] 沈杰, 乔少杰, 韩楠, 等. 融合多信息的个性化推荐模型[J]. 重庆理工大学学报(自然科学), 2021, 35(3): 128-138.

[10] 崔春生, 杜柏瀚, 王雪. 基于分层序列的移动电子商务推荐系统策略研究[J]. 数学的实践与认识, 2020, 50(8): 12-21.

[11] 黄玲, 余霞. 基于云平台的电子商务商品智能推荐系统[J]. 现代电子技术, 2020, 43(5): 183-186.

[12] 丁昭巧. 多 Agent 技术下电子商务用户感知和个性化推荐模型探究[J]. 商业经济研究, 2018(3): 98-102.

[13] 郑祥云, 陈志刚, 黄瑞, 等. 基于 SP_LDA 模型的商品推荐算法[J]. 小型微型计算机系统, 2016, 37(3): 454-458.

[14] 李艳琦. 信息中介参与电子商务系统的系统动力学分析[J]. 工业技术经济, 2016, 35(10): 30-37.

[15] 吕仲明. 基于深度强化学习的列表推荐系统[D]: [硕士学位论文]. 烟台: 烟台大学, 2024.

[16] 黄甲. 基于 Multi-Agent 与本体的个性化旅游推荐系统建模研究[D]: [硕士学位论文]. 大连: 东北财经大学, 2018.

[17] 姚楠, 何山, 赵越, 等. 结合强化学习和用户短期行为的新闻推荐算法[J]. 计算机应用与软件, 2024, 41(4): 284-290.

[18] 董相宏, 安俊秀. 基于动态动作覆盖的深度强化学习新闻推荐[J]. 大数据, 2024, 10(3): 109-118.

[19] 黄雄. 个性化推荐算法在电子商务系统的实践探索[J]. 信息与电脑(理论版), 2024, 36(10): 165-167.

[20] 胡子豪. 基于图神经网络的多行为推荐系统研究[D]: [硕士学位论文]. 南昌: 华东交通大学, 2023.