计算机科学
Computer Science
ISSN 1002-137X,CN 50-1075/TP

《计算机科学》网络首发论文

题目：基于强化学习的推荐研究综述
作者：余力，杜启翰，岳博妍，向君瑶，徐冠宇，冷友方
收稿日期：2021-02-08
网络首发日期：2021-06-28
引用格式：余力，杜启翰，岳博妍，向君瑶，徐冠宇，冷友方．基于强化学习的推荐研究综述．计算机科学.
https://kns.cnki.net/kcms/detail/50.1075.TP.20210628.1621.008.html

基于强化学习的推荐研究综述

余力[1] 杜启翰[1] 岳博妍[1] 向君瑶[1] 徐冠宇[2] 冷友方[1]

1 中国人民大学信息学院 北京 100872
2 北京理工大学徐特立学院 北京 100081

摘 要 推荐系统致力于从海量数据中为用户寻找并自动推荐出有价值的信息和服务，可有效解决信息过载问题，成为大数据时代一种重要的信息技术。但推荐系统的数据稀疏性、冷启动和可解释性等问题，仍是制约推荐系统广泛应用的关键技术难点。强化学习是一种交互学习技术，该方法通过与用户交互并获得反馈来实时捕捉其兴趣漂移，从而动态地建模用户偏好，可以较好地解决传统推荐系统面临的经典关键问题，强化学习已成为近年来推荐系统领域的研究热点。文中从综述的角度，首先在简要回顾推荐系统和强化学习的基础上，分析了强化学习对推荐系统的提升思路，对近年来基于强化学习的推荐研究进行了总体梳理与总结，并分别从传统强化学习推荐和深度强化学习推荐的研究情况进行总结；在此基础上，重点总结了近年来强化学习推荐研究的若干前沿，以及其应用研究情况。最后，对强化学习在推荐系统中应用的未来发展趋势进行分析与展望。

关键词：推荐系统；强化学习；深度强化学习；马尔可夫决策过程；多臂老虎机

中图法分类号 TP183 DOI 10.11896/jsjkx.210200085

Survey of Reinforcement Learning Based Recommender Systems

YU Li[1], DU Qi-han[1], YUE Bo-yan[1], XIANG Jun-yao[1], XU Guan-yu[2] and LENG You-fang[1]

1 School of Information, Renmin University of China, Beijing 100872, China
2 XUTELI School, Beijing Institute of Technology, Beijing 100081, China

Abstract Recommender systems are devoted to finding and automatically recommending valuable information and services for users from massive data, which can effectively solve the information overload problem, and become an important information technology in the era of big data. However, the problems of data sparsity, cold start, and interpretability are still the key technical difficulties that limit the wide application of the recommender systems. Reinforcement learning is an interactive learning technique, which can dynamically model user preferences by interacting with users and obtaining feedback to capture their interest drift in real time, and can better solve the classical key issues faced by traditional recommender systems. Nowadays, reinforcement learning has become a hot research topic in the field of recommendation systems. From the perspective of survey, this paper first analyzes the improvement ideas of reinforcement learning for recommender systems based on a brief review of recommender systems and reinforcement learning. Then, the paper makes a general overview and summary of reinforcement learning based recommender systems in recent years, and further summarizes the research situation of traditional reinforcement learning based recommendation and deep reinforcement learning based recommendation respectively. Furthermore, the paper summarizes the frontiers of reinforcement learning based recommendation research topic in recent years and its application. Finally, the future development trend and application of reinforcement learning in recommender systems are analyzed.

Keywords Recommender systems; Reinforcement learning; Deep reinforcement learning; Markov decision process; Multiple arm bandits

到稿日期：2021-02-08 返修日期：2021-05-21
基金项目：国家自然科学基金（71271209）；中国人民大学研究基金（2020030228）
This work was supported by the National Natural Science Foundation of China (71271209) and Research Foundation of Renmin University of China (2020030228).
通信作者：余力（buaayuli@ruc.edu.cn）

1 引言

近年来，随着互联网、大数据、云计算等信息技术的迅猛发展，人们被暴露在规模日益增长的大体量数据环境中[1]。大数据中蕴含着丰富的信息与知识，使得人们可以在短时间内获取大量信息。但同时，信息爆炸带来的负面影响就是导致用户面对海量信息时，难以从纷繁复杂的数据中获取到真正有价值的内容，从而面临“信息过载”问题。推荐系统作为一种信息过滤技术，通过向用户提供个性化的内容来解决信息过载问题，成为诸多应用领域的关键技术[1]，也是目前学术界关注的热点前沿。

从技术上讲，最主要的推荐技术主要有两类，即基于协同过滤的推荐[2]和基于内容的推荐[3]。前者依赖于用户间的影响关系，后者主要从内容特征属性计算匹配程度。近年来，通过深度学习（Deep Learning），系统可以深层次学习和表征用户和项目的非线性特征，从而提取到用户和项目的本质特征[4]，成为众多学者关注的焦点。尽管深度学习技术极大地促进了用户和项目的潜在特征学习，推动了推荐系统在学术研究和应用深度，但是，在许多场景下，难以获取足够的用户偏好信息，稀疏性冷启动等问题仍然是制约推荐系统的关键瓶颈[5]。同时，传统的推荐方法在刻画用户的偏好时，把每个项目看作是独立的，无法刻画其序列特征，不能对序列中的项目关系建模。

近年来，强化学习方法在游戏、机器人控制等领域取得了突破性的进展[6-7]，已经成为人工智能时代新的研究热潮，同时也给推荐领域的研究带来了新的机遇。结合深度学习方法之后的强化学习具备了处理大规模数据、发现并提取底层特征的能力，从而更准确地实现特定的目标。作为一种交互式推荐（Interactive Recommendation, IR）方法，基于强化学习的推荐模型可以通过与用户进行实时交互、获得用户真实反馈来更新推荐策略，相较于传统静态方法更符合现实推荐场景[8]；同时，由于强化学习问题通常被规范化为马尔可夫决策过程（Markov Decision Process, MDP），因此这类模型具有建模用户行为序列[7]的天然特性，可以充分刻画序列特征并捕捉用户的动态偏好；并且，其中的探索机制的设置可以使智能体更加充分地探索状态、动作空间，在一定程度上提高了推荐结果的多样性；最后，由于这类模型常以最大化推荐系统的累计收益，也就是用户的长期反馈作为优化目标来更新推荐策略，可以在一定程度上提高用户的长期满意度[9]。

笔者采用 Google Scholar 中的高级搜索功能，以关键词“Reinforcement Learning & Recommendation”以及“Reinforcement Learning & Recommender Systems”进行搜索，得到近 5 年来强化学习与推荐系统相结合的学术文献出版情况，如图 1 所示。可以看出，近年来与基于强化学习的推荐系统相关的研究论文不断增多。

从目前来看，强化学习在推荐系统中的研究应用已经成为近年来的研究焦点之一。然而，目前还缺乏相应的综述总结，Zhao 等[6]对深度强化学习在搜索、推荐系统、广告领域的应用进行了综述，但是其在推荐系统相关研究方面的篇幅较小。除此之外，据我们所知，目前还没有中文文献较为系统化地以推荐系统为研究对象，并对基于强化学习[10]的实现方法进行全面系统的总结和分析。

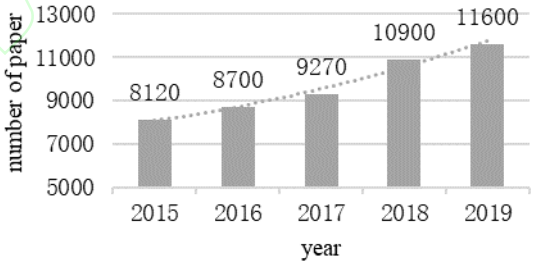

图 1 近 5 年基于强化学习的推荐系统文献数量

Fig. 1 Paper quantity of reinforcement learning based recommender systems in the last 5 years.

本文主要对基于强化学习的推荐系统的研究与应用进展进行综述。第 2 节简要介绍了推荐系统与强化学习技术，分析了推荐系统中目前存在的关键问题，简要介绍主要的强化学习方法；第 3 节总体介绍了基于强化学习的推荐研究体系框架，并重点介绍了基于传统强化学习的推荐；第 4 节重点介绍基于深度强化学习的推荐；第 5 节重点讨论了现有基于强化学习的推荐的若干前沿研究；第 6 节介绍了主要的强化学习推荐应用领域；最后做出总结与研究展望。本文的整体结构思路如图 2 所示。

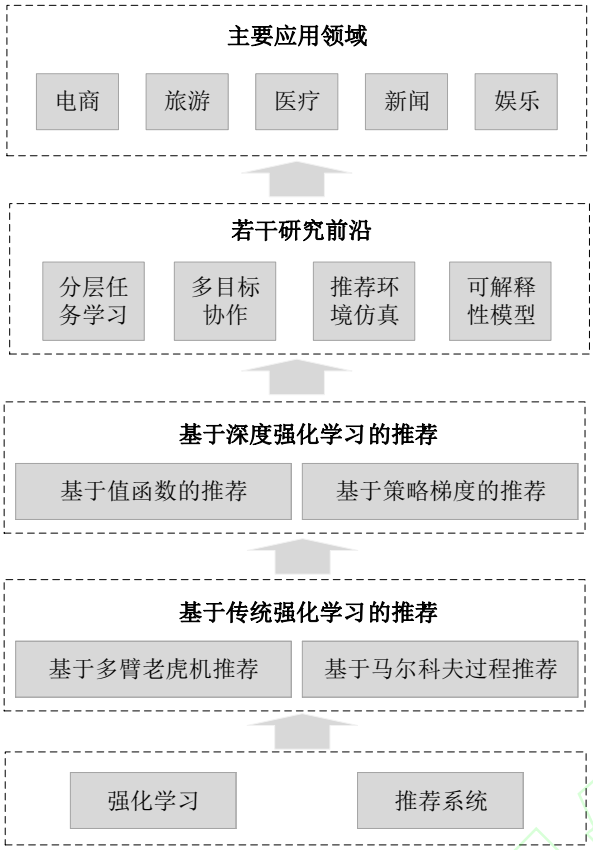

图 2 文章整体框架

Fig. 2 Overall framework of the paper

2 推荐系统与强化学习综述

本节主要介绍了推荐系统和强化学习技术，以及二者在深度学习背景下的发展，并分析了推荐系统的现存问题。

2.1 推荐系统

2.1.1 推荐系统简介

典型的推荐系统任务一般定义为在推荐环境中为用户生成推荐项目列表或预测用户对某个特定项目的喜爱程度，项目的类型包括商品、服务、信息等。目前，推荐算法应用在诸多领域，例如电子商务推荐（阿里巴巴、京东、ebay 等）、数据检索推荐（百度、谷歌、搜狗等）、服务推荐（美团、大众点评、网易云音乐）、新闻推荐（今日头条、腾讯新闻）等。

从技术来讲，推荐方法通常被分为 3 类：基于协同过滤的推荐、基于内容的推荐、混合方法推荐。基于协同过滤的推荐系统基于用户-项目的历史交互记录[11]产生推荐，

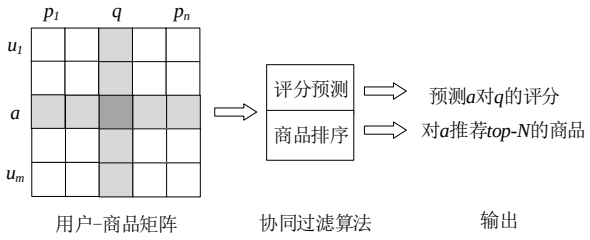

图 3 协同过滤算法示例[5]

Fig. 3 Example of collaborative filtering algorithm [5]

可以是显性反馈（评分、喜欢/不喜欢），也可以是隐性反馈（浏览、点击），如图 3 所示。基于内容的推荐主要根据用户和项目的特征信息来进行推荐[12]，首先建立用户偏好向量和项目属性向量，然后计算两者之间的相似度来为用户推荐相似度最高的项目。混合方法推荐结合了以上两种不同的推荐方法，即可同时建模静态特征与动态交互。

2.1.2 基于深度学习的推荐

深度学习是当今非常重要的智能技术。借助深度学习技术，我们可以建立包含多个非线性模块的深度模型来挖掘用户的潜在偏好，从而学习到项目（商品）的本质特征和主题。因此，深度学习近年来在推荐系统领域受到了广泛关注。一般地，基于深度学习的推荐框架[4]包含 3 层：输入层、模型层、输出层。其中，输入层主要以用户画像（性别、年龄、喜好等）、项目内容（类别、文本、图像等）、用户反馈（点击、评分、浏览等）和辅助信息（标注、评论等）中的一种或多种组合为输入；模型层由各种深度学习模型堆叠、线性或非线性组合而成，以学习用户和项目的潜在表示；输出层通过计算相似度等方法从项目池中召回部分项目，并通过个性化排序算法产生最终推荐结果[4]。目前，推荐系统领域常用的深度学习模型包括多层感知机（Multi-Layer Perception，MLP）、自编码器（Auto-Encoder，AE）、循环神经网络（Recurrent Neural Network，RNN）、卷积神经网络（Convolutional Neural Network，CNN）、注意力机制（Attention Mechanism，AM）对抗生成网络（Generative Adversarial Network，GAN）等，这些模型在不同场景下的推荐任务上展现出了优越的性能。基于深度学习的推荐系统框架如图 4 所示。

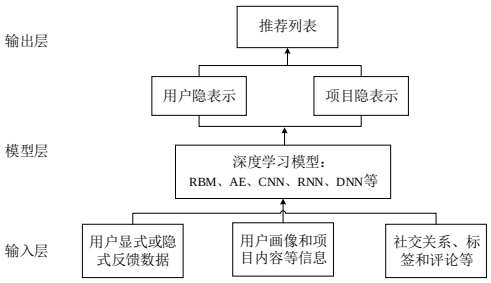

图 4 基于深度学习的推荐系统框架

Fig. 4 Framework of deep learning-based recommender systems

2.1.3 推荐系统研究现存问题

基于上述分析，目前推荐系统存在以下4个关键问题。

（1）用户长短期兴趣建模与冷启动

一般地，以用户最近交互的 K 个物品序列来建模短期兴趣，以用户 ID 与属性信息的嵌入向量来建模长期兴趣。虽然基于 CNN / RNN / Transformer 网络的长短期兴趣建模方式得到业界广泛认可[13]，但是在冷启动条件下难以合理建模新用户的长期兴趣，可能构建错误的用户画像。

（2）多目标联合优化

在基于用户会话（Session）与场景（Scene）的推荐方法上，需要实现多目标优化来平衡评价指标。在模型结构上，子任务根据相关性共享子网络参数[14]，但是高相关性的子任务往往会占用大部分权重，实际上仅优化了少数目标[15]；在权重分配上，采用帕累托最优（Pareto Optimality）进行权重组合寻优[16]的方法取得了一定成果，但是帕累托最优解的不唯一性表明这种方法并不稳健。

（3）特征提取与多模态信息融合

为了使推荐系统完成信息互补并理解用户行为，需要融合多模态信息作为新的特征输入。例如，采用 ResNet 或 ReceptionNet 提取图片特征[17]的方法取得了显著效果，但是深层网络结构极大地限制了特征提取速度，因此迫切地需要一种通用的特征提取器来取代 DeepFM 等 DNN 网络，使其能够快速高效地融合推荐系统所需要的多模态信息。

（4）推荐系统的可解释性

未来的推荐系统希望成为一个用户透明的、用户参与的交互系统，所以必须使推荐结果可以被用户理解。基于知识图谱的推荐系统[18]通过构建<实体,关系,属性>三元组大大提高了可解释性，但是知识图谱的大数据规模难以根据用户反馈进行实时更新，并且很难提前过滤脏数据。目前，知识图谱大多作为边界信息（Side Information）来略微提高推荐的可解释性，仍然缺少一类可以进行实时交互更新的可解释性推荐方法。

2.2 强化学习技术

强化学习技术的本质是通过最大化即时奖励与未来奖励的折扣和来达到最优目标[18]。它是机器学习中一种用来解决序贯决策的重要方法，其采取持续的“交互-试错”机制，通过与环境的不断交互学得最优的有效策略。广义来讲，强化学习技术主要包括 3 个方面，分别是多臂老虎机（MAB）[19]、马尔可夫决策过程（MDP）[20]、深度强化学习（DRL）。本节对这些技术展开讨论，并且在最后汇总了与本综述密切相关的强化学习最新前沿发展。

2.2.1 传统强化学习技术

（1）多臂老虎机

多臂老虎机问题（Multi-armed Bandit, MAB）定义为[21]：给定 K 个可能的行为，每个行为都与固定未知的奖励概率分布相关；在每次迭代中，智能体选择一个行为进行游戏并获得奖励，其任务是学习如何选择单次行为以随时间的推移最大化累积收益。本质上，MAB 问题是探索（Exploration）与开发（Exploitation）的平衡问题（EE 问题）[22]。Bandit 算法是解决 EE 问题的一种有效途径。其目标是最小化累积遗憾函数，经典的 Bandit 算法有：Epsilon-Greedy 算法、Thompson sampling 算法[23]、UCB 算法[22]等。与原始的多臂老虎机方法不同，上下文赌博机算法（Contextual Multi-Armed Bandit, CMAB）考虑到了场景上下文[24]，引入了状态的概念。智能体使用状态描述来采取更明智的行动。相较于单一赌博机，上下文赌博机情境中有多个赌博机，环境的状态告诉智能体它此时正在对付哪一台赌博机。智能体的目标是学习针对任意数量的

赌博机的最优行动。上下文 Bandit 算法有 LINUCB[25]、Neural Bandit[26]、上下文 Thompson Sampling[27]等。

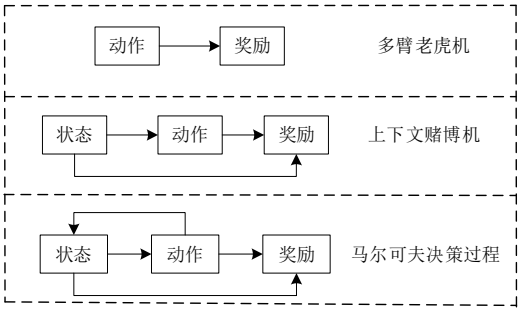

图 5 多臂老虎机、上下文赌博机和马尔可夫决策过程之间的区别

Fig. 5 Difference between Bandits, Context Bandits and MDP

（2）马尔可夫决策过程

马尔可夫决策过程（MDP）指状态间存在依赖且存在时间概念的序贯决策过程，其涉及 5 个元素：环境状态s、行动a、奖励R、状态转移概率矩阵$\boldsymbol{P}^{\boldsymbol{\pi}}_{\boldsymbol{s_t,s_{t+1}}}$和折扣系数$\gamma$。MDP 中当前时刻状态仅与上一时刻状态有关。基于 MDP 的强化学习方法把整个推荐过程看作一个马尔可夫决策过程[20]，智能体通过一定的策略π依据状态转移概率矩阵$\boldsymbol{P}^{\boldsymbol{\pi}}_{\boldsymbol{s_t,s_{t+1}}}$选择一个行为与环境产生交互，交互产生后环境会给智能体一个反馈奖励R。其中，个体在给定状态s_t下从行为集合中选取一个行为a_t的依据称为策略π，它是基于行为集合的一个概率分布。多臂老虎机与马尔可夫决策过程之间的区别如图 5 所示。强化学习的目的是通过不断地调整策略来最大化长期累计奖励，其通过定义两种函数来衡量每个状态或状态-动作对的好坏，分别是状态价值函数$V_\pi(s)$和行为价值函数$Q_\pi(s,a)$，最优策略π^*则为可以使得状态价值函数和行为价值函数取最大值的策略。基于马尔可夫决策过程的强化学习方法有 Q-learning，Sarasa 等。

2.2.2 深度强化学习技术

深度强化学习（Deep Reinforcement Learning, DRL）将强化学习与深度学习结合起来，打破了基于马尔可夫决策过程强化学习的存储限制。根据深度神经网络的逼近目标不同，我们可以将 DRL 分为两类：基于值函数的 DRL 和基于策略梯度的 DRL[27]。

基于值函数的 DRL：即 DQN（Deep Q Network）及其变体。DQN 将 Q-learning 算法与深度学习结合在一起使其可以应对较大的状态空间[28]。其本质是将 Q 函数表用神经网络近似代替，这样就可以避免表存储结构的有限性。DQN 算法包括 Double-DQN[29]，Dueling-DQN[30]，Prioritized Replay Buffer[31]，Categorical DQN，NoisyNet[32]，Distributional DQN[33]，Rainbow[34]等。DQN 是基于 Q-learning 来确定损失函数，目标是使目标 Q 值和估计 Q 值之间的差值越小越好。基于策略梯度的 DRL：基于值函数的学习算法虽然能够高效地解决连续状态空间的强化学习问题，然而其行为空间依然是离散的；同时，在面对观测受限、随机策略的学习时该方法也不再适用。此时基于策略梯度（Policy Gradient）的学习便是解决这类问题的一个途径。策略梯度算法输入状态，直接输出动作或动作的概率；训练时通过奖励值来判断一个动作的相对好坏[35]；但是其参数必须要经历一个完整的回合才可以进行更新。行动者评论家（Actor-Critic）方法是一种联合基于价值函数和策略函数的算法，策略函数充当行动者（Actor）生成行为与环境交互，行为价值函数充当评论家（Critic），对行动者的行为进行评价并指导它向更优的方向更新。DDPG（Deep Deterministic Policy Gradient）算法[35]在 Actor-Critic 框架的基础上加入双网络结构以及经验回放机制来更有效地学到最优的策略。

2.2.3 强化学习前言研究技术

目前的强化学习前沿研究方向有分层强化学习、多任务迁移强化学习、多智能体强化学习、基于记忆和推理的强化学习等[7]。分层强化学习方法[36]可以将相对复杂困难的整个大任务分解成一个个规模较小的子任务以促进其高效学习；多任务迁移强化学习的目标是训练一个独立的模型可以用于多个不同任务中；多智能体强化学习模型已经被用于一些存在合作或者竞争关系的难题中；基于记忆和推理的强化学习目前实质上是通过加入外部的记忆组

件，使得智能体具有了初步的主动认知和推理能力。

3 基于强化学习的推荐系统总体框架

本节首先分析了强化学习对推荐系统的提升思路；然后对当前主要的强化学习推荐研究进行了系统的分类，并重点介绍目前在传统强化学习方面的研究。

3.1 强化学习对推荐系统的提升思路

通过对强化学习技术的分析，强化学习能够对推荐系统的以下方面产生提升作用。

（1）实时获取用户动态偏好

基于强化学习方法的推荐是一种交互性推荐方法。交互性推荐系统（Interactive Recommender Systems）在个性化服务中扮演了很重要的角色[8]。传统方法是静态方法，不符合推荐中与用户动态交互的场景。相较于传统的静态推荐方法来说，交互性推荐在向用户推荐商品之后，接收用户给予反馈，再针对用户此刻的反馈调整推荐策略。因此其推荐策略是根据用户的实时反馈及时调整的。

（2）捕捉推荐项目间的关联关系

与会话型推荐相同，强化学习方法通过对用户点击序列建模，捕捉序列间的项目关系。而其他传统的推荐方法没有考虑序列关系，不能很好地捕捉到用户的动态偏好变化[37]。用户的兴趣是动态变化的，一个用户的兴趣会随着时间、年龄的变化而不断改变，这体现在用户的点击序列中。例如，用户刚开始喜欢娱乐新闻，但是随着时间迁移慢慢更倾向于浏览政治新闻。

（3）探索机制避免重复推荐

传统推荐方法会推荐大量重复相似项目给用户，大大降低了用户效用[37]。传统推荐方法通过挖掘出用户的历史偏好以推荐相似项目，因此会推荐大量相似项目。而强化学习方法中的探索机制可以巧妙地避免向用户推荐大量重复商品的问题，给用户带来惊喜，提高推荐的准确度[38]。

（4）关注用户长期满意度

传统推荐方法旨在提高用户即时满意度，如点击率等，而忽略长期满意度（如用户持续使用时长）[37]。强化学习方法是通过最大化即时奖励和未来奖励的折扣和来动态更新策略的，是以提高用户长时满意度为目标[7]。因此强化学习方法能提高用户留存率以及单次使用时长等反映长期满意度的评估指标。

3.2 基于强化学习的推荐研究分类

目前基于强化学习的推荐研究依赖于对标准的强化学习模型进行应用拓展。总体来看，我们将其分为两个大类，分别是基于传统强化学习的推荐和基于深度强化学习的推荐。其中传统强化学习细分为两类，分别为基于多臂老虎机的推荐和基于马尔可夫决策过程的推荐。基于深度强化学习的推荐细分为基于值函数 DRL 的推荐和基于策略梯度 DRL 的推荐。

3.3 基于多臂老虎机的强化推荐

基于多臂老虎机（MAB）的方法主要使用各种各样的 Bandit 算法来做推荐，该方法的出发点是平衡探索-利用（Exploration-Exploitation）之间的关系，不仅能推荐给用户之前喜欢的商品的相似商品，还能创新性地探索用户其他的偏好以避免重复性推荐，给用户带来更多的惊喜。

（1）改进现有 Bandit 算法进行推荐。对于不同的推荐任务，学者们开发出改进的 Bandit 算法进行推荐。Wang 等[39]将探索-利用之间的平衡作为强化学习任务，使用贝叶斯模型了解用户的偏好，同时考虑了音频的内容和推荐的创新性，使用线性逼近和变分推理算法加速贝叶斯推理；Wu 等[40]提出了一种协同上下文 Bandit 算法，该算法利用用户之间的邻接图共享上下文和相邻用户之间的收益，同时进行在线更新，与传统的独立 Bandit 算法相比，其严格地证明了所提出的协作 Bandit 算法是一个改进的后悔上界；Broden 等[41]提出了一种汤普森抽样 Bandit 策略（Thompson Sampling）的扩展，用于编排电子商务基本

推荐算法的集合，关注商品对商品的推荐问题，为集成学习者提供了多个基于行为和属性的预测器，并且展示了当行动可用性和奖励的平稳性都没有得到保证时，如何使汤普森抽样适应现实情况；Wang 等[42]通过基于因子分解的 Bandit 算法进行在线交互推荐，低秩矩阵完成是在一个增量构造的 user-item 偏好矩阵上执行的，在这个矩阵上，开发了一个基于上置信界的项目选择策略，以平衡在线学习期间的开发/探索权衡。利用用户之间可见的上下文特征和依赖关系（如社会影响）来提高算法的收敛速度，帮助克服推荐中的冷启动问题。Intayoad 等[43]提出一种相关性敏感的上下文 Bnadit 算法来解决在线课程推荐问题，与专注于推荐列表新颖性的方法不同，其考虑到用户普遍存在的重复学习行为，将历史学生行为（Past Student Behavior, PSB）和当前学生状态（Current Student State, CSS）以成对的方式进行建模，并建立相关性矩阵来量化候选动作（即候选的课程）之间的相关性，以最大化用户的累积点击次数作为优化，从而提高在线学习平台的用户粘度。

（2）引入深度神经记忆模块以减少人机交互。Shen 等[44]考虑到现有的强化学习解决方案需要与每个用户进行大量交互才能提供高质量的个性化建议，为了减轻这种限制，他们设计了一种新的深度神经记忆增强机制，根据每个用户以前的交互来为其建模和跟踪历史状态。因此，用户对新项目的偏好可以通过少量的交互快速了解。

3.4 基于马尔可夫决策的强化推荐

基于马尔可夫决策过程（MDP）的强化学习是研究的比较早的领域，虽然越来越多的学者已经投入到深度强化学习的研究中，但对于一些状态空间和动作空间较小或者离散的问题来说，采用马尔可夫决策建模可以降低推荐的时间复杂度，因此仍有较多基于 MDP 的推荐研究。

（1）对于状态动作空间较小的推荐问题研究。Zhang 等[45]将多智能体强化学习（Multi-agent RL）方法运用到学者的动态合作者推荐问题中，根据多个相似度衡量指标计算两两之间的相似度，把社交网络中的每个学者的影响力值即该学者与其他所有学者的相似度总和作为一个状态，状态的 status 值是通过该学者的原始值加上其邻居学者中最大的 status 的折扣值更新，最后向其推荐相似度最高且影响力最大的学者；Liebman 等[46]将基于马尔可夫过程的强化学习方法用于音乐歌单推荐中，他们把为用户推荐歌单的过程看作马尔可夫过程。通过模型训练为每个用户规划出一个满意度最高的歌单。歌单推荐的奖励由两部分定义构成，

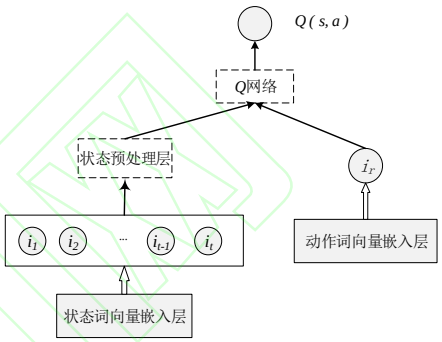

图 6 基于值迭代 DRL 的推荐框架

Fig. 6 The framework of value-based deep reinforcement learning for recommendation

即用户对每首歌曲的评分以及歌单上的歌曲序列转移，后者作为另外一种奖励。

（2）减少系统状态动作空间以提高可扩展性。大多数推荐背景中动作状态空间巨大，单纯将马尔可夫决策过程应用于推荐中时，可学习到的策略有限，可扩展性低。因此有学者通过限制系统的状态动作空间来解决这个问题。Choi 等[47]采用双聚类技术，将推荐系统设计成一个网格世界游戏，极大地减少了系统的状态和动作空间。采用双聚类技术不仅减少了空间，而且有效地提高了推荐质量，解决了冷启动问题，同时可以为项目推荐提供解释。Frits 等[48]考虑到推荐对可用容量的影响，在多智能体约束的部分可观测决策问题中，利用一种新的信念空间采样算法，通过限制后悔来限制状态空间的大小。通过利用问题的平稳结构，该算法比现有的近似求解器具有更强的可扩展性。

4 基于深度强化学习的推荐系统

深度强化学习是近几年深度学习与强化学习相结合的方法，得益于深度学习近年来的快速发展，深度强化学习的推荐也成为近年来学者关注的焦点。本节从模型及其要素、训练与奖励函数设计方面重点介绍了基于深度强化学习推荐的具体解决方案。

4.1 模型框架研究

4.1.1 基于值函数的强化推荐

基于值函数 DRL 的推荐使用深度神经网络来近似模拟 Q 值函数，以最大化总奖励为优化目标，通过梯度下降不断更新神经网络参数以找到最优策略。

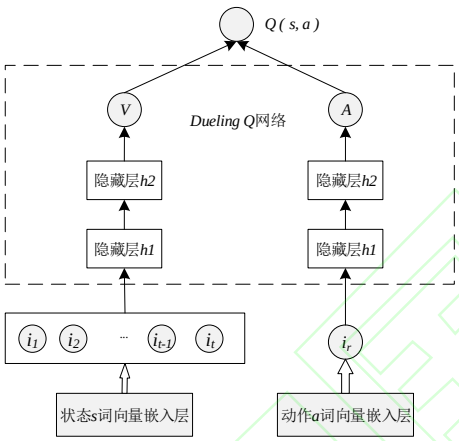

图 7 DRN[38]模型示例

Fig. 7 Example of DRN[38] model

本文把现有的用于推荐系统的基于值函数 DRL 框架总结为图 6，而不同的学者对框架中不同的部分进行改进以达到他们特有的目的。图 6 中实线框表示必有的层，虚线框表示不同的作者添加或者优化的不同层，对于每一层，不同的学者可能会给出不同的定义。模型输入是状态 S 和动作 A，之后通过嵌入层把输入转化成低维向量。这里有学者把状态做了转换，比如使用 RNN 提取状态的时间序列关系，对状态和动作进行关系提取之后再输入到 Q 网络中输出状态-动作对的 Q 值。不同的文章使用了不同的 Q 网络（如 DQN，Double DQN，Dueling DQN 等）。

（1）将 Dueling-DQN 用于用户动态推荐。在一些推荐情境下，用户的动作选择只与当前状态有关，所以在 Nature DQN 模型的基础上，Dueling-DQN[30]将 Q 网络分为两部分，分别计算状态值函数$V(S)$以及依赖状态的动作优势函数$A(s,a)$，以此来挖掘单纯的状态变化对用户决策的影响。Zheng 等[37]提出的 DRN 采用 Dueling-Double-DQN 网络结构来捕捉用户的新闻偏好随着新闻变化的动态性。其中状态价值函数可以提取仅由状态决定的奖励。他们将用户特征和上下文特征用来表示当前的状态，将新闻特征和新闻用户交互特征用来表示当前的一个动作。这些特征经过模型可以输出当前状态采取这个动作的预测 Q 值。同时模型采用 Dueling 结构，该模型架构如图 7 所示，与图 6 相比 Zheng 等主要对 Q 网络进行了修改。

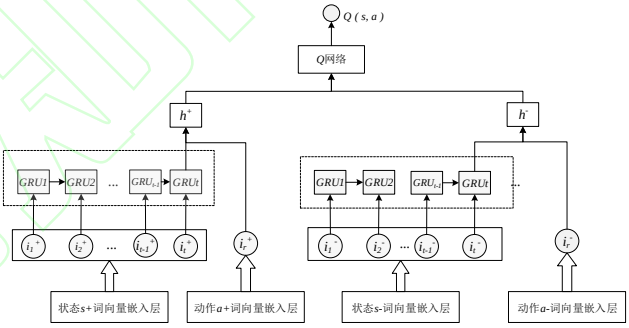

图 8 DEERS[9]模型示例

Fig. 8 Example of DEERS[9] model

（2）考虑正负反馈的 DQN。基于强化学习的推荐是通过学习用户的反馈来得到最优策略选择，反馈包括正反馈和负反馈，负反馈如用户忽略或跳过这个商品，代表用户对这个商品不感兴趣。但之前的算法都没有考虑到负反馈，在用户跳过该商品时状态不变。这样就会导致系统继续推荐相似的商品。因此 Zhao 等[9]提出的 DEERS 模型考虑加入负反馈作为奖励的一部分，以避免向用户推荐其不喜欢的商品。正负反馈均会更新相应的状态，但是负反馈的数量明显多于正反馈，因此把它们结合起来学习是具有挑战性的，考虑到这个难题，他们提出了 DEERS 推荐框架来解决。图 8 为 DEERS 的模型框架示意图，他们把 ***s+*** 和推荐的商品 ***a*** 拼接起来作为正向信号，把 ***s-***和推荐的商

品 ***a*** 拼接起来作为负向信号。***s*+**定义为用户最近点击的商品序列，***s*-**定义为用户最近忽略的商品序列。

（3）以优化用户长短期满意度为目标的 DQN。现有的推荐系统一般都是把提高商品的点击通过率（CTR）作为目标，而没有考虑推荐结果对用户的长期满意度（如推荐公平性[49]、会话时长等）的影响。因此，有研究在推荐时考虑到对用户长短期满意度的协同优化。为了优化推荐列表中商品展示顺序，Liu 等[50]将两种监督学习信号加入奖励函数，其中分类信号监督商品的正负样本标签，而排序信号监督成对商品的展示顺序，以此提高推荐列表质量；Chen 等[51]将用户行为的个性化约束加入奖励函数，降低对频繁行为的奖励，提高对新行为的奖励，鼓励用户对推荐系统进行长期探索；Zou 等[52]提出基于 DQN 的 FeedRec 模型，其创新性地提出优化用户的长期满意度，可以表现在用户浏览的持续时间、用户活跃度等。为了能够有效地反映用户的延时反馈，Q-network 被重新设计为 3 层：Raw behavior

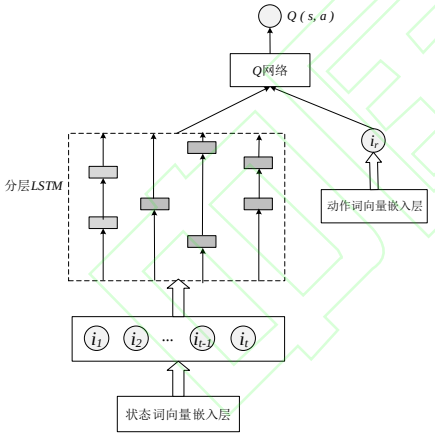

图 9 FeedRec[52] 模型示例

Fig. 9 Example of FeedRec[52]model

embedding layer，Hierarchical Behavior Layer，Q-value Layer。Raw behavior embedding layer 的作用是输入原始的用户行为信息提取出用户的状态，以为之后的优化做准备。用户的行为包括忽略、点击、购买，每个行为代表的含义各不相同，因此为了准确捕捉不同的行为，Zou 等应用分层的 LSTM。如图 9 所示，整个模型图相较于图 6 而言主要是在状态预处理加了一层 LSTM 层。

Chang 等[53]提出的 value-aware 推荐算法使用值迭代思想最大化期望商品交易总额（Gross Merchandise Volume, GMV）来实现推荐系统的最大经济效益，用户的点击、加入购物车和购买等行为与商品的被点击率、价格等信息集成为 GMV。在该算法的离线训练上，Chang 等以真实历史数据的 NDCG 为标准构建外部奖励函数，激励智能体将用户实际点击和购买的商品排在推荐列表前面。这种较为简洁的排序方法在准确率和召回率上均优于基于深度神经网络的排序。

（4）基于值函数 DRL 的多目标组合优化推荐。在推荐时系统往往是推荐一个列表或者一页的序列，而当直接优化整个推荐页的奖励时，强化学习方法的探索复杂度很高。基于此，Eugene 等[54]提出了基于 Q-learning 的 SlateQ 算法。其将一个页的推荐序列分解成多个项目，分别计算每个项目的长期收益 LTV（Long-term Value）。 然后将各个项目的长期收益加入排序多目标中进行优化。同时，对于 Q 值的多目标组合优化问题，他们提出了基于 LP（Linear Program）的优化以及两个启发式算法，并在推荐环境仿真器 RecSim 上验证了 SlateQ 的鲁棒性。

（5）推荐背景下改进的 Dyna-Q 框架。Dyna-Q 框架集成了线上学习与线下规划，可以提高训练效率。强化学习应用到推荐中需要与真实用户交互，而这样会降低用户满意度。因此有学者通过构造环境仿真器模拟用户行为，让智能体从历史行为中习得推荐策略。然而在使用历史数据构造仿真器时,现有研究都没有考虑到历史数据选择的有偏性，并且仿真器都是在推荐策略学习之前构造的，而且始终保持不变。Zou 等[55]提出 Pseudo Dyna-Q（PDQ）框架中构造的World模拟环境会随着策略的更新而不断更新，以使得模拟更接近于真实情况。同时，他们提出的 Pseudo Dyna-Q 是根据历史数据与 World 仿真器的数据共同作用来提升策略的。使用 World 仿真器的数据规划有效避免了 off-policy 学习的 Deadly Triad 问题。

（6）结合协同过滤的 DQN 模型。Lei 等[56]提出的 UDQN（User-specific Deep Q-Network）是一种结合矩阵

分解与 DQN 的推荐算法，将经矩阵分解预训练好的用户、项目向量作为嵌入向量，通过向用户顺序推荐商品并获得反馈的方式更新状态向量，模拟真实交互中用户偏好的变化过程；同时，由于用户嵌入与项目嵌入的点积即为评分，可以利用这种固有的监督信号来规约状态的更新方向。Lei 等[57]进一步提出的 GCQN（Graph Convolutional Q-Network）使用图卷积神经网络来聚合用户或项目的一阶邻居节点，从而建立了融合用户-项目二部图中的图结构化信息的高维状态表示，随后进一步采用 GRU 建模状态更新过程，并将 GRU 在每个时刻输出的隐藏状态向量经全局注意力机制加权聚合，再与候选项目向量拼接输入全连接网络来预测该状态-动作对的 Q 值。上述方法分别利用矩阵分解和图神经网络来整合潜在协同信息，提高了 DQN 的预测准确率与稳定性。

4.1.2 基于深度策略梯度算法的推荐

（1）将深度策略梯度算法作为推荐系统的一个模块。Zhang 等[58]使用策略迭代算法来优化监督学习的模型参数，模型主体是监督学习，而没有直接使用强化学习方法来进行推荐。Zhao 等[59]使用策略梯度算法更新对抗生成网络框架中的生成器模型参数，解决了生成推荐列表任务中离散采样无法直接使用梯度下降的问题。Sun 等[60]把深度策略梯度算法用于对话系统中，使用 Policy-based 方法来决定对话机器人在某一时刻是否开始进行推荐。

（2）策略修正的深度策略梯度方法。在现实企业中应用强化学习方法推荐时，Agent 在学习更新新的策略时用到的往往是基于之前的策略得到的历史数据（用户的轨迹），而不是在现有策略基础上得到的轨迹，这会导致策略学习产生很大的偏差。这时 Agent 学到的是旧策略喜欢推荐给用户的商品。基于此，Chen 等[61]提出了 off-policy 的策略修正方法。该方法是在 Policy-Gradient 算法 REINFORCE 基础上进行的改进。其在 YouTube 上进行的实验表明此方法有效。

（3）针对大规模动作空间中交互性推荐问题的优化。基于 DQN 和 DDPG 的强化学习算法都需要从所有项目中挑取一个可以使奖励最大化的项目，而推荐系统中的项目数量庞大，导致这个过程的时间复杂度很高。基于此，Chen 等[62]针对大动作空间中的交互性推荐问题，提出基于树结构的策略梯度模型 TPGR。他们提出在项目上建立一个平衡的层次聚类树，将项目的选择归结为从树的根到某一叶的路径选择，大大降低了训练和决策阶段的时间复杂度；然后结合策略梯度模型来做出决策。为了训练和评估模型，他们设计了一个环境模拟器对标准公开数据集中的用户行为进行模拟，大量的实验表明 TPGR 模型效果优异。

（4）基于深度策略梯度的上下文赌博机推荐方法。上下文赌博机方法可以保证探索和利用之间的权衡以及最小化在线成本，因此被广泛应用于决策支持中。然而，目前的上下文赌博机方法往往过度简化问题的假设，因此其在真实业务场景中的适用性有限。Pan 等[63]提出了基于策略

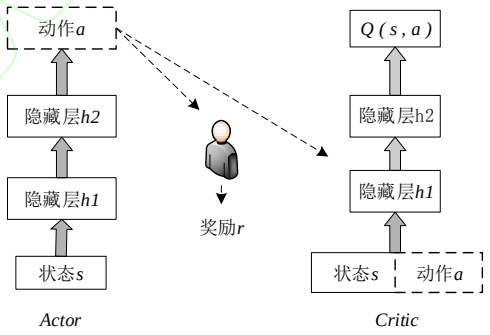

图 10 基于行动者评论家 DRL 的推荐框架

Fig. 10 The framework of Actor-Critic based deep reinforcement learning for recommendation

梯度的上下文推荐方法 PGCR。它在没有不切实际的假设下仍然可以处理问题，PCGR 把搜索空间限制到特定的策略中，因此可以大大减少策略选择的空间。同时，其使用了经验回放机制，不仅可以解决上下文赌博机问题，还能解决马尔可夫决策过程等问题。

（5）基于模型（model-based）的深度策略梯度推荐算法。现有的研究大都是不基于模型的（model-free），这种方法中策略的训练学习需要智能体与真实环境进行大

量的频繁交互，学习代价高昂。Bai 等[64]提出了一个基于模型（model-based）的 DRL 解决方案 IRecGAN。它通过一个生成对抗网络，对用户与智能体间的交互建模以为离线的策略学习提供支持。IRecGAN 使用鉴别器来评估生成数据的质量并对生成的奖励进行缩放，这样同时减少了所学到的用户模型和策略的偏差。

4.1.3 基于行动者评论家 DRL 的推荐

基于行动者评论家 DRL 的推荐方法的基本原理是使用 Actor 来训练策略输出动作，Critic 来评价这个动作的好坏，然后 Actor 再根据 Critic 的评价进行策略调整。我们将行动者评论家算法在推荐系统中的基本框架总结为图 10。

（1）基于 AC 算法和监督学习的推荐。SRL-RNN[65]是基于大规模电子健康记录的动态医疗方案推荐系统。之前的医疗方案推荐系统均使用单一监督学习方式（通过匹配指标信号和医生处方）或者单一强化学习方式（通过最大化表明存活率的评估指标）。但是没有研究考虑到结合这两

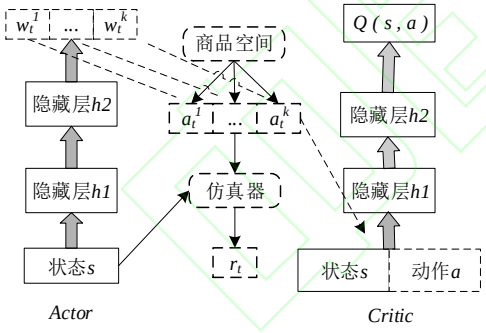

图 11 LIRD[63]模型示例

Fig. 11 Example of LIRDF[63] model

种方式的优点。因此 SRL-RNN[65]创造性地结合了监督学习与强化学习。SRL-RNN 框架应用 actor-critic 框架来处理多种药物、疾病和个人特征之间的复杂关系；框架中的 actor 是由指标信号和评估信号两者共同调整其行为，确保了有效的处方和低死亡率；该框架进一步使用 RNN 来解决由于在现实生活中可能无法观察到全部状态而导致的部分观测马尔可夫过程（POMDP），即缺失部分状态值的马尔可夫过程。SRL-RNN 的整个框架包括 3 个主要的网络：Actor（Actor-target），Critic（Critic-target），LSTM。Actor 网络根据患者的动态状态来推荐随着时间变换的医疗方案，而医生的处方提供的指标信号保证了安全性以及加速了学习过程。Critic 网络评估与 Critic 网络相关的行为价值，以鼓励或阻止推荐的治疗。

（2）基于 AC 算法的商品清单推荐。Zhao 等[66]提出对电商平台的用户进行 List-wise 的推荐，即推荐一个购物清单，但与以往 list 不同的是，一个清单上物品之间是有互补关系的，而不是简单的 Top *N* 的推荐，这可以避免推荐很多相似的商品。奖励在推荐系统推广到线上之前是很难得到的，所以他们做了一个线上的环境仿真器用来预训练参数，仿真器是基于历史数据的，考虑到历史数据中并不包含所有的 state 和 action，其基于没见过的状态-动作对p_t与历史状态-动作对m_i的相似性来计算其奖励。LIRD 模型采用 Actor-Critic 框架，整个框架如图 11 所示。

（3）基于 AC 算法的二维购物清单页面推荐。Zhao 等[67]提出对电商平台的用户进行 Page-wise 的推荐，即推荐

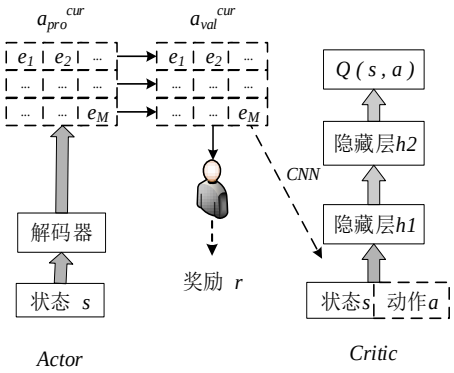

图 12 DeepPage[65]模型示例

Fig. 12 Example of DeepPage[65] model

一个基于页面的二维购物清单，DeepPage 能捕捉到用户对于二维页面上商品位置的视觉偏好。DeepPage 使用很多不同的网络结构如 RNN，CNN 来更加准确地捕捉用户的点击序列偏好以及页面视觉偏好。Actor 网络使用了

encoder-decoder 结构来编码状态空间，解码动作空间；Critic 则使用 DQN 来做值函数近似，DeepPage 的模型结构如图 12 所示。

（4）基于多模态信息的强化学习推荐方法。用户的简单反馈如浏览、点击、评分等往往不能完全体现出用户真正的偏好点。而用户文本信息反馈能够充分反映用户真实的偏好所在，同时商品视觉语义信息反映了用户的视觉偏好。Zhang 等[68]提出了基于文本信息的强化学习推荐方法。推荐系统很容易违反用户过去通过文本评价反馈的偏好信息。为了解决这个问题，他们提出一种全新的约束增强的强化学习架构 RCR，它能够随着时间的推移有效地合并用户偏好。具体地，他们利用一个鉴别器来检测推荐系统是否做出违反用户历史偏好的推荐，这会被加到强化学习的奖励函数中，从而实现最大化期望的累积未来奖励。Yu 等[69]提出的推荐框架 VL-Rec 将商品视觉语义与用户评论文本融入智能体的奖励函数中，具体地，actor 以目标商品视觉语义、属性与用户偏好之间的最小欧氏距离为特征更新策略，critic 的奖励函数与用户评论文本确保了用户历史偏好的顺序更新。

4.2 模型要素定义

在基于深度强化学习的推荐系统中，输入数据往往是用户现在所处的环境状态、采取的动作，而对应不同的应用和研究需要给出针对性的定义。

（1）考虑上下文特征的新闻推荐模型要素定义。Zheng 等[37]提出的 DRN 模型是针对新闻的推荐，其每次输入有 4 个特征，分别为新闻特征、用户特征、新闻和用户的交互特征、上下文特征。新闻特征包括题目、作者、排名、类别等共 417 维；用户特征包括用户在 1h、6h、24h、1 周、1 年内点击过的新闻的特征表示，共 2065 维；上下文特征包括 32 维的时间、日期和新闻实效性等。将用户特征和上下文特征用于表示当前的状态（state），新闻特征和交互特征用于表示当前的一个动作（action）。

（2）商品列表推荐中模型要素定义。Zhao 等[66]提出的 LIRD 模型是典型的电商平台中的推荐，其推荐任务是推荐一个商品清单，商品之间具有互补关系。LIRD 将状态定义为用户点击或购买的最新的 N 个商品。动作定义为要推荐给用户的商品列表。状态转移定义为：当前的状态（state）是用户最近浏览的 N 个物品。动作（action）是新推荐给用户的 K 个商品，如果用户忽略了全部的这些商品，那么下一个时刻的状态（state）和当前的状态是一样的，如果用户点击了其中的两个物品，那么下一个时刻的 state 是在当前 state 的基础上，从前面剔除两个商品同时将点击的这两个物品放在最后得到的。

（3）考虑负反馈的推荐模型要素定义。Zhao 等[9]提出的 DEERS 是针对电商系统中的推荐，他们考虑到负反馈的重要性，因此将状态 $\boldsymbol{s}$ 定义为：$\mathbf{s} = (\boldsymbol{s}_+, \boldsymbol{s}_-) \in \mathrm{S}$，即既有正反馈的状态也有负反馈的状态，其中 $\boldsymbol{s}_+ = \{i_1, \ldots, i_N\}$为用户最近点击的或者购买的 N 个商品，$\boldsymbol{s}_- = \{j_1, \ldots, j_N\}$定义为用户最近忽略的 N 个商品，以时间顺序排列。状态转移 s 到 s'则定义为：当推荐系统在状态 $\boldsymbol{s}$ 时推荐商品 $\boldsymbol{a}$ 给用户，如果用户忽略了这个商品，那么 $\boldsymbol{s}'_+ = \boldsymbol{s}_+$并且更新$\boldsymbol{s}'_- = \{j_2, \ldots, j_{N,} a\}$，如果用户点击或者购买了这个商品，更新$\boldsymbol{s}'_+ = \{i_2, \ldots, i_N, a\}$同时$s'_- = s_-$。同时他们利用 GRU 网络来捕捉用户最近点击的 N 个商品$\{i_1, \ldots, i_N\}$之间的序列关系，

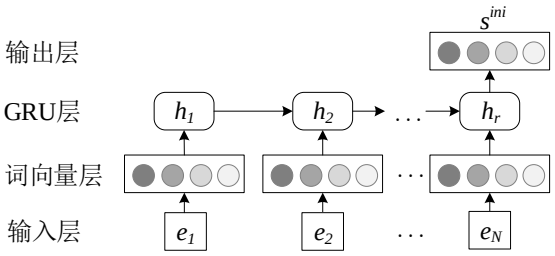

图 13 初始状态编码器[65]示例

Fig. 13 Example of initial state encoder [65]

并将最后一个隐藏层的输出 h_N 作为状态s_+； s_-同理。

（4）考虑用户长期满意度的模型要素设计。Zou 等[52]提出的 FeedRec 是针对电商系统的推荐，它考虑到用户长期满意度的重要性，因此将状态定义为：以用户属性表示

初始状态，随后根据用户的异构行为来差异化更新状态。考虑到用户的行为包括忽略、点击、购买，每个行为代表的含义各不相同，比如忽略代表用户对该商品不感兴趣，点击代表用户对该商品有兴趣，而购买代表用户对该商品极其感兴趣，为了准确捕捉不同的行为，他们在 Q-network 中加入了分层的 LSTM。

（5）知识图谱增强的推荐模型要素设计。Wang 等[70]提出的融合知识图（Knowledge Graph, KG）信息的强化学习推荐模型的改进是：1）状态s_t由 3 个部件拼接而成，h_t是用户交互的历史项目序列经 GRU 建模的高维表示，c_t是经 KG 嵌入的项目知识图的高维特征，f_t是经知识推理网络预测的用户偏好；2）奖励函数的两个部件是R_{seq}与R_{kg}，分别衡量序列质量与知识图质量。Chen 等[71]提出了一种首先使用注意力机制聚合 KG 的状态向量，再在用户-项目兴趣图中通过图卷积操作获取 Q 值。实验证实了加入先验知识能有效提升智能体的推荐性能。

（6）二维购物清单页面推荐模型要素设计。Zhao 等[67]提出的 DeepPage 虽然也是应用于电商平台的推荐模型，但它面对的基于 page-wise 的推荐任务要求向用户推荐一个二维的购物清单页。DeepPage 使用了 encoder-decoder 来编码状态空间，解码动作空间。他们使用 GRU 网络来捕捉用户的系列初始偏好，输入为用户最近点击/购买的商品序列$\{e_1,\dots,e_n\}$，输出为用 GRU 最后一层隐藏状态$\boldsymbol{h_t}$表示的用户初始偏好向量$\boldsymbol{s^{ini}}$。这种初始状态编码器结构如图 13 所示。。

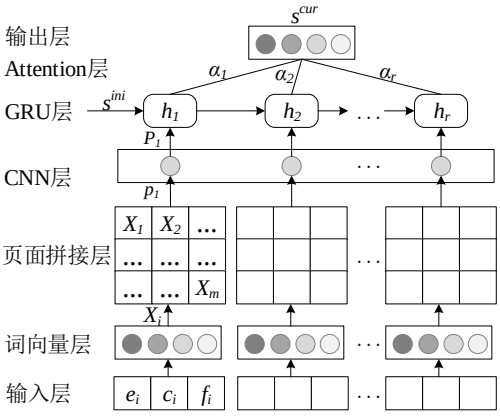

图 14 实时状态编码器 [65]示例

Fig. 14 Example of the real-time state encoder [65]

实时状态编码空间的产生如图 14 所示。网络的输入为当前页面中的商品表示以及用户的反馈$\{X_1,\dots,X_M\}$，其中，$\boldsymbol{x_i}=(\boldsymbol{e_i},\boldsymbol{c_i},\boldsymbol{f_i})$，$e_i$为商品特征表示，$\boldsymbol{c_i}$为商品分类（one-hot 编码），$\boldsymbol{f_i}$为用户反馈（one-hot 编码），$\boldsymbol{x_i}$经过嵌入层转化成低维度向量。再把这些低维向量按照原始的页面顺序拼接起来，然后用一个 2D-CNN 来进一步捕捉商品展示之间的位置关系。最后用一个 GRU 捕捉用户的序列偏好。同时他们使用之前输出的$\boldsymbol{s^{ini}}$作为 GRU 的初始状态。并且，为了捕捉用户在当前 session 的实时偏好，他们应用了注意力机制。其中，GRU 的长度是根据当前 session 的长度调整的。对状态向量执行反卷积操作得到的输出即为动作向量，公式如式（1）所示。文献[67]Critic 使用 DQN 来做函数值近似。同时，由于解码产生的动作向量可能并不对应真实的商品向量，因此需要根据相似度寻找最近邻的商品进行推荐。

$$a^{cur}=deconv2d(s^{cur}) \tag{1}$$

（7）使用注意力机制和记忆网络表征状态。Zou 等[55]提出的 Pseudo dyna-q 框架中状态跟踪模块使用注意力机制和记忆网络来捕捉状态轨迹的长时序列关系。注意力机制为不同时间的项目设置了不同的权重，记忆网络相较于循环神经网络来说可以捕捉到更久远项目的特征。

从目前的调研来看，很多学者都使用循环神经网络、注意力机制、记忆网络、卷积神经网络等深度学习方法来

提取低维的序列关系。总的来说，对于输入数据进行精准的定义和处理会直接影响到模型的学习效果，因此这一环是学者们需要首先注意的研究要点。

4.3 模型训练设计

深度强化学习是强化学习与神经网络的结合，而神经网络的训练需要的输入数据是独立同分布的，强化学习数据具有马尔可夫性，数据之间关联非独立。如果直接把关联数据用于训练会导致神经网络参数无法收敛，因此有学者提出经验回放机制，通过对经验池中的数据进行随机采样而得到相互独立的数据再进行训练。但实际上，利用这种方法对网络进行训练时仍然存在收敛不稳定的情况。因此对于训练的设计是至关重要的一环。

（1）优先经验回放机制和分离的双网络结构。Zhao 等[66]提出的 LIRD 框架使用 Actor-Critic 结构，因此在训练时他们采用 DDPG 算法同时使用优先经验回放机制和分离的双网络结构（target-net 和 eval-net），来减小收敛不稳定性的问题，同时设计了一个仿真器来模拟用户的行为选择，模型在线下用仿真器来进行预训练可以减少用户的探索成本。

（2）环境仿真器辅助训练。Zou 等[52]考虑到在推荐系统上线之前，需要先对模型使用历史数据进行 off-policy 训练，以避免上线的模型盲目地进行试错从而大幅降低用户满意度，而且由于历史数据可能并不包含上线后环境中不同的状态-动作对，导致策略学习的收敛不稳定，因此设计了一个 S 网络来模拟真实环境中用户的反馈，以此来辅助 off-policy 的策略学习。Zou 等[55]通过建立一个 World Model 来模拟推荐环境，其会随着策略的更新而不断更新，以使得模拟更接近于真实情况。

（3）用户模型辅助训练。Zhao 等[72]提出的 DeepChain 框架使用 Actor-Critic 结构，训练时同样使用 DDPG 算法。与其他很多研究不同的是，DeepChain 模型是基于模型（model-based）的强化学习算法，需要提前获知环境的状态转移概率，Zhao 等开发了一个概率网络（probability network），使用监督学习方法来做非线性的函数值近似，以预测相应的状态转移概率。Chen 等[73]使用 model-based 的强化学习模型进行推荐，他们使用对抗生成网络来模拟用户行为选择函数以及奖励函数，然后再结合 Q 网络训练，这种对抗生成用户模型效果优异。

4.4 奖励函数设计

对于狭义的强化学习方法来说，奖励（reward）作为马尔可夫决策过程中的一大元素，它的定义对于算法的效果起着至关重要的作用，因为它是表示用户对推荐结果满意度的指标，而我们的目标设定就是最大化用户满意度即最大化奖励。因此有很多学者对于推荐中不同场景的奖励进行了更为精确的定义。

（1）奖励的补充

在电商推荐系统中，奖励一般定义为用户对于该推荐商品的反馈（忽略、点击、购买），只有产生点击或者购买时，才会更新状态，但其实这样定义并不是最优的，比如用户的忽略行为往往远多于点击行为，购买行为则更少，无法捕捉负反馈会导致用户偏好学习产生偏差[9]。或者大多数系统只是简单地把点击通过率（CTR）作为奖励，却忽略了直接长期奖励如用户活跃度等。

1）考虑负反馈的奖励函数。Zhao 等[9]考虑到负反馈对于反映用户偏好的重要性，提出深度强化学习框架 DEERS，把负反馈也作为一种奖励以避免向用户推荐其不喜欢的商品，正负反馈均会更新相应的状态，将正负反馈分割开来作为 DQN 的输入，其中正反馈是指用户点击了系统推荐的商品，负反馈是指用户跳过了系统推荐的商品。Zhao 等[9]在京东的数据集上对该框架进行了实验，证明了其有效性。Gao 等[74]同样考虑将负反馈作为奖励之一，其使用 CNN 模型来捕获顺序特征以获得正反馈；然后，通过对抗训练学习最佳的负反馈表示，将正/负表示同时输入到 DQN 中，这有助于生成更好的动作值函数。

2）考虑用户长期满意度的奖励函数。Zheng 等[37]提

出的 DRN 模型考虑到了用户的活跃度也是反映其对推荐内容的喜好表示，用户活跃度体现在他进入系统的频率。他们使用了 DQN 网络来捕捉建模新闻的动态变化属性，把用户活跃度作为奖励的补充以更好地获取用户偏好。Zou 等[52]考虑到现有的推荐系统一般是把提高商品的点击通过率（CTR）作为目标而忽略了用户长期满意度的优化，长期满意度可以表现在用户浏览的持续时间、用户活跃度等，因此他们创新性地提出优化用户的延时反馈矩阵，采用了一个包含分层 LSTM 网络的 Q 网络 FeedRec，其定义的奖励函数使得模型可以同时优化即时反馈和延时反馈。奖励使用 3 种矩阵按一定权重配比来表示，分别为：点击矩阵（click metric）、浏览深度矩阵（depth metric）以及活跃度矩阵（return time metric）。其中点击矩阵为即时反馈，浏览深度矩阵和活跃度矩阵为延时反馈。Chang 等[53]将商品交易总额（Gross Merchandise Volume ,GMV）的期望作为奖励函数，GMV 由用户的点击、收藏和购买等行为与商品的被点击率、价格等信息集合而成。

（2）考虑奖励的差异化

对于不同的用户以及不同的环境状态，奖励应该是有差异的，因此在推荐环境下设计奖励函数时应该考虑到这种差异。

Chen 等[51]提出的 Robust DQN 模型把深度强化学习用于电商推荐系统，他们考虑到目前很多深度强化学习方法都是应用于比较稳定的环境中，比如 AlphaGo 围棋零和博弈的环境是固定的，状态也是确定的，这时的奖励能很准确地表明真正的奖励。而对于电商推荐系统来说，环境是动态变化的，比如双十一时会产生大量的交易，此时的奖励并不是推荐系统促进的而是环境的变化影响的，因此奖励的估计就会产生很大的误差。为了解决这个问题，他们引入了近似后悔奖励，通过参考小样本的用户数据，消除环境变化对奖励估计带来的影响。同时在不同的时间里用户的分布是变化的，使得奖励的方差很大，因此他们提出使用分层抽样经验池回放机制。该机制是针对电商系统中的提示（tip）推荐，总奖励由 3 部分组成：1）r_1由是否点击该 tip 以及点击后的驻留时间组成，公式如式(2)所示，若点击则$I=1$，否则$I=0$，e是用户点击 tip 的总次数，x是当前的页面号，ρ是反映不同用户偏好的个性化系数。。2）r_2反映了用户使用 tip 机制的频率， y 是用户最近的100个页面浏览中点击tip的次数，公式如式(3)所示。3）r_3表示用户是否根据该 tip 成功购买了商品，若是则$c=1$，否则$c=0$，公式如式(4)所示。总奖励由以上 3 部分线性加权组成，公式如式(5)所示。

$$r_1 = I^* (1+\rho \quad ^*e - \lambda \tag{2}$$

$$r_2 = I * e - y \tag{3}$$

$$r_3 = c \tag{4}$$

$$r = r_1 + \alpha * r_2 + \beta * r_3 \tag{5}$$

Chen 等[73]提出使用对抗生成网络来模拟奖励函数及用户选择行为函数，然后使用这两个函数线下训练 model-based Q 网络，训练完备后可直接在线调整策略。

总而言之，上述研究从两个方面对奖励的设置进行了优化：1）对奖励的定义进行补充，如加上负反馈或者用户长期满意度指标；2）消除环境变化和用户差异对奖励估计带来的影响。但是目前的研究都是通过手工定义奖励的，并不能为用户定制个性化的奖励函数。比如在电商系统中用户的行为有忽略、点击、加入购物车、购买等，而不同的用户有不同的行为偏好，可能用户 A 对一个商品极其感兴趣时才会加入购物车，而用户 B 对一个商品稍微有一点感兴趣就会加入购物车。这时手工定义的奖励并不准确。因此有学者使用对抗生成网络来模拟奖励函数[75][76]，实验结果表明，这种方式比手工定义奖励更加准确。

5 强化学习推荐的若干前沿进展

近年来，基于强化学习推荐的逐步纵向深度，传统研究更注重于强化学习模型本身的应用。近几年，有较多学者从不同的视角对强化学习推荐的前沿问题进行研究，包

括分层强化学习、多智能体强化学习；同时，很多学者将强化学习应用于前沿的推荐领域中，如可解释性推荐、社交化推荐等。本节试图梳理若干重要前沿进展。

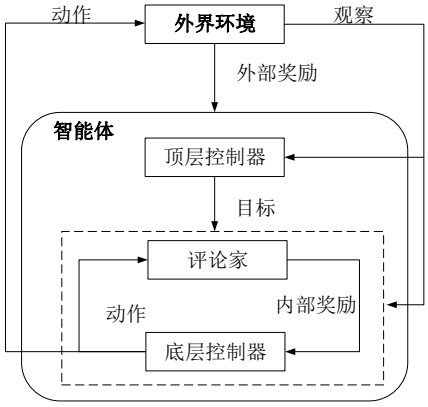

图 15 分层深度 Q 网络模型结构图[7]

Fig. 15 Framework of hierarchical deep Q network model [7]

5.1 基于分层强化学习的推荐

分层强化学习方法通过将最终目标分解为多个层的子任务来学习层次化的策略，并通过组合多个子任务的策略形成有效的全局策略[7]。在复杂的推荐任务中，直接以最终目标为导向来优化策略的效率较低，此时利用分层的方法将复杂任务拆解可以有效地提高推荐效率。分层的 DQN 模型结构如图 15 所示。

目前大多推荐任务以提高推荐项目点击率为目标，很少考虑到推荐项目的转化率。实际上，转化率指标更能体现用户的真正偏好，而由于转化率数据极其稀疏不易学习，因此目前大多模型在推荐转化率上表现得并不好。当反馈很稀疏时，智能体不能进行有效的学习，因为极少的反馈使得智能体在一些重要状态空间的探索不够充分。

基于时空抽象和内在激励的分层 DRL 算法可以通过把总目标分解成抽象子目标来降低学习复杂度，其可以在存在稀疏反馈问题的 DRL 任务中保持高效的探索。Zhao 等[77]提出了基于分层强化学习的推荐模型 MaHRL。其中，高层智能体专注于学习用户长期偏好，以购买序列与点击序列为输入，从而输出多个抽象目标来反映用户兴趣，并指导低层智能体决策；底层智能体以点击序列和浏览序列为输入以学习用户的短期偏好，可直接产生推荐动作来满足抽象目标。Xie 等[78]提出了一种分层模型来完成集成推荐任务，高层智能体负责在特定频道（如视频、图文）下根据用户细粒度偏好推荐项目，而低层智能体则根据用户的粗粒度偏好选择频道。这些分层目标设置对应着分层奖励函数，通过增加奖励信号降低了模型的学习困难问题。

在使用注意力模型在进行推荐时，给历史项目序列中用户不感兴趣的项目（噪音项目）赋权会分散用户喜爱项目的权重，从而最终使得目标推荐项目的权重甚至低于随机推荐项目。基于此，Zhang 等[79]将分层强化学习用于辅助注意力推荐模型的权重修正中。他们提出采用分层强化学习的方法来剔除历史课程序列中的噪音课程。其中高层智能体负责判断用户的整个历史序列是否需要剔除课程，如果需要剔除，则低层智能体判断历史序列中的每一个课程是否需要剔除；如果不需要剔除，则低层智能体不采取行为。通过分层的学习，把任务分解成两层子任务，从而降低了任务的复杂性。

5.2 基于多智能体强化学习的推荐

基于多智能体强化学习的方法通过多个智能体的策略学习进行协作优化，适用于一些需要竞争合作的推荐任务场景，此时基于单智能体的强化学习方法难以适应。例如，在电商平台中一个完整的用户会话（session）在历经首页、商品详情页或购物车页面时都会得到推荐列表，这是由不同场景下各自独立的排序策略所生成的。根据博弈论理论的古诺双寡头模型（Cournot Duopoly Model）可知，独立优化排序策略虽然会短暂地提高某个场景的收益，但是最终可能会降低电商平台的期望总收益。因此，我们将讨论如何进行协作优化以获得最大期望总收益。

在电商推荐任务中，用户在登录电商平台后首先会进入首页，其中包含多个系统推荐的商品，用户会选择点进商品详情页或者忽略接着浏览首页，如果用户进入商品详情页则会有新的推荐商品。而在现有的系统中这两种场景下的

推荐却是不相关的。Feng 等[80]将不同推荐场景下的推荐组合成一个目标。其提出的 MA-DPG 利用多个推荐智能体来分别完成不同场景下商品推荐任务，，由整个会话中唯一的评价网络集中评估所有场景的行为，由于智能体只能借助通讯模块观察局部环境来优化独立策略，这就迫使多智能体通过协作的方式来最大化唯一评价网络的奖励。但是，MA-DPG 作为一种 Model-free 的强化学习算法，对训练数据规模存在较高的要求，并且算法的收敛并不稳定。因此，He 等[81]基于相关均衡理论设计了一种熵正则化的信号网络来协调多智能体间的奖励信号，从而提高算法收敛到全局最优策略的速度。Zhao 等[72]把电商平台上的首页、商品详情页的两个推荐场景结合成一个推荐目标，即在一种场景中需考虑到另一种场景中客户的行为,进而提出了 DeepChain 模型。他们提出基于 whole-chain 的推荐，即在用户建立的整个会话（session）中不同场景的推荐都有考虑到用户此会话的历史行为。他们设计了两个不同的智能体分别做决策，分别是首页和商品详情页中的行动者（Actor），而只有一个评论家（Critic）同时调整两个智能体的行为。DeepChain 是基于模型的强化学习方法（model-based），相较于无模型的算法不需要大量的用户与智能体间的交互。其模型示例如图 16 所示。

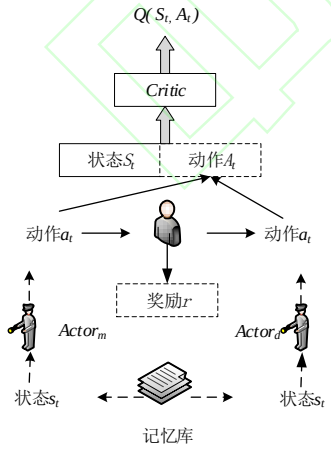

图 16 DeepChain[72]模型示例

Fig. 16 Example of DeepChain[72] model

Gui 等[82]把多智能体强化学习用于社交平台发文时@的人选推荐任务中。用户在推特上发文时@的人选不仅与该用户的历史推特信息密切相关，同时也与潜在参与用户的偏好息息相关。因此在推荐人选时，不仅需要习得该用户的历史偏好，也需要参考其他用户的历史推特信息。因此，他们使用两个智能体来协同学习目标用户的偏好。一个智能体用来选出目标用户的所有历史相关推特，同时另一个智能体用来选出潜在参与用户的历史相关推特。通过两个智能体的合作选出所有最相关的历史推特，以进一步习得用户的偏好。

5.3 面向深度强化推荐环境仿真器研究

基于强化学习的推荐方法需要在与现实的环境交互中训练和评估，而将强化学习方法直接放到线上进行训练成本很高。因此，研究者们通过构造一个环境仿真器对环境进行模拟，并使用环境仿真器进行线下训练，这种方法可以大大降低训练成本。同时，对于序列推荐的研究者来说，通用的仿真平台对于其算法和模型更深入的研究也具有深远的意义。

Shi 等[83]通过淘宝用户真实的历史购买记录构造了一个虚拟淘宝平台，以为后续的强化学习推荐、排序等任务提供训练环境支持。他们提出 GAN-SD 算法模拟虚拟买家的操作和搜索请求操作，并且提出了多智能体对抗模仿学习（MAIL，Multi-agent Adversarial Imitation Learning），同时学习买家规则和平台规则，以训练虚拟用户与平台产生更加真实的交互。为了缓解训练中的过拟合问题，他们提出了动作规范约束 ANC 策略（Action Norm Constraint）。实验结果表明，与传统的监督学习方法加静态数据的推荐方法相比，使用他们提出的虚拟平台与强化学习推荐方法结合产生的推荐策略，效果总是更优。

MAIL 框架假设只存在两个相互交互学习的智能体，即用户智能体把平台当作环境，平台智能体把用户当作环境。而在真实的应用环境中，用户的决策不仅会受到平台的影响，往往还会受到外界各种各样的干扰，而这些干扰在数据中无法体现。基于此，Shang 等[84]在对环境进行重

构时考虑到了干扰因素的影响，把潜在干扰因素作为另外一个智能体，与另外两个智能体产生交互，进而提出了 DEMER（deconfounded multi-agent environment reconstruction）模型。DEMER 借助模仿学习和对抗生成网络的方法来重构环境以达到仿真的效果。Shang 等在滴滴出行真实场景下对 DEMER 进行了验证，结果表明其相较于已有算法有显著的性能提升。

在模仿用户行为时存在很多挑战，如潜在的项目分布是复杂的、每个用户的历史记录是有限的。为了解决这些难题，Zhao 等[85]使用对抗生成网络生成模拟行为。其中生成器根据历史日志生成新的日志，判别器不仅要辨别出真假日志，还要预测用户的行为。

静态的数据无法反映真实情况下用户与系统之间的交互情况，也无法预测新策略下用户的行为，因此使用静态数据对强化学习方法评测也是困难的。为了推进基于强化学习的交互性推荐任务的研究，有研究人员开发出通用的仿真平台提供支持。David 等[86]构建了一个商品广告推荐环境的模拟器，以为强化学习任务提供支持。Shi 等[87]构建了一个可以支持各种标准推荐数据集（如 movielens，Yelp 等）的环境模拟器 PyRecGym。OpenAIGym 实现的是各种游戏的环境模拟器，而 PyRecGym 与 OpenAIGym 类似，是针对基于强化学习的推荐应用设计的推荐环境模拟。其主要定义了 3 个函数，分别是 Initialization，Reset，Step。谷歌的研究人员 Eugene 等[88]谷歌的研究人员开发了一套可以配置和自定义的仿真平台 RecSim，用于为强化学习与推荐特别是协同交互性推荐的研究提供支持。RecSim 环境中包含 3 个部分，分别是用户模型、文件模型以及用户选择模型。其通过向用户推荐一系列资料或列表来与环境发生交互，并通过模拟个体用户对于文件的观测特征进行推荐。特别地，RecSim 允许研究者自定义仿真用户的行为，并提高了那些使用 RecSim 平台进行仿真训练的模型学习用户异构行为序列的能力。

5.4 基于强化学习的可解释性推荐

推荐系统在人们日常生活中的影响越来越深入，因此如何让用户更加理解和相信系统的推荐结果越来越重要。可解释推荐在给用户推荐项目的同时解释推荐原因，使得用户更加信任推荐结果。

很多推荐机制是复杂且难以解释的，此时需要在对推荐结果进行事后解释。即把推荐模型与解释模型分离开，用单独的模型推荐结果做出解释。Wang 等[89]使用强化学习方法对推荐结果生成解释。在其提出的可解释框架中，被解释的推荐模型作为环境的一部分，对强化学习方法生成的句子解释进行奖励或惩罚。框架中有两个智能体与环境交互，第一个智能体根据当前状态生成句子解释，第二个智能体根据第一个智能体生成的句子解释来预测用户对所有商品的评分。如果此评分与被解释的推荐模型的预测评分相似则被奖励。同时如果智能体给出的句子解释满足可读性、连贯性高、解释精炼等条件，智能体也会被给予奖励。通过这两个奖励条件更新两个智能体的策略，不仅可以使其习得解释能力，并且也保证了事后解释的质量。

Xian 等[90]提出了一种基于知识图谱与强化学习推理的可解释性推荐方法。知识图谱中包含丰富的用户、项目信息，可以对推荐的解释性问题提供直观有力的信息支持。然而要把知识图谱中的用户-项目节点对之间的所有路径都枚举出以进行相似性计算难以实现。因此他们通过训练用于搜索路径的智能体，把强化学习方法用于解释性推荐中。以知识图谱为环境，智能体在训练阶段要学习的策略是从用户导航到潜在的感兴趣项。如果达到正确的项目，智能体会得到环境给予的较高奖励。因此，在策略训练收敛后，智能体可直接遍历正确的推荐项目，而不必枚举用户-项目对之间的所有路径，路径为项目推荐提供解释。

McInerney 等[91]为可解释性推荐开发了一个多臂老虎机方法。他们认为不同的用户对解释信息的反映各不相同并且是动态变化的。因此，他们提出的基于多臂老虎机的探索-开发平衡方法的目的是为每个用户找到最佳的解释序列。此方法不仅可以学习到每个用户对于哪些解释信息

做出了何种反映，并且也可以学习到对于每个用户来说哪些项目是推荐的最佳项目，以及如何在探索与开发之间取得平衡以应对不确定性。实验表明，解释信息会影响用户对推荐内容的反映。这项工作表明，多臂老虎机方法中的探索-开发方法不仅有利于推荐任务，也提高了推荐的可解释性。

5.5 其他前沿研究

考虑社交网络影响的推荐。用户在选择商品时受到周围朋友的偏好影响[92]。因此在对用户推荐商品时不仅要考虑用户本身的个人偏好，还要考虑到周边人对其决策的影响。Lei 等[92]提出了社交注意力 DQN 模型 SADQN，即在应用 DQN 进行推荐的同时考虑到其社交网络中亲密朋友对用户的偏好影响。其 Q 值由两部分组成，分别是代表个人偏好的函数 Q^p 和代表社交偏好的函数 Q^s。其社交偏好由一层社交注意力层计算得到。

端到端的基于 DRL 的推荐。现有的基于强化学习的推荐研究基本上都由 3 部分组成，分别是向量嵌入表示部分（Embedding Component，EC）、状态表示部分（State Representation Component，SRC）和策略学习部分（Policy Component，PC）。而现有的研究中 EC 部分都是通过预训练得到的，且在后续的状态表示和策略学习中固定不变。用户和项目的关系是动态变化的，而预训练得到固定不变的嵌入向量不能很好地表示用户或项目。由此，Liu 等[93]提出端对端的 RL-based 推荐方法 EDRR，使得 EC 部分可以与 SRC 和 PC 部分协同训练。为了避免 EC 部分训练过程中的不稳定性，他们引入了一个监督学习的信号。其提出了 3 个引入监督学习的途径。实验表明，EDRR-v3 框架在基于值函数 DRL 和基于策略梯度 DRL 的方法中均可以实现端到端的推荐，并且表现最稳定。

6 强化学习在推荐系统中应用研究

由于强化学习可以动态获取用户的行为信息，实时融入最新的偏好信息，目前越来越多被地应用于新闻、电商、医疗等领域。

（1）电商领域

电商领域一直是推荐研究最重点关注的领域，也是强化学习推荐的重点。有较多研究采用基于 MDP 的强化学习方法把用户和推荐系统之间的序列化交互看作一个马尔可夫过程，通过强化学习技术来自动学习最优的推荐策略。Zhao 等[66]提出基于 List-wise 的推荐，这样更能提供给用户多样性的选择；Zhao 等[67]提出了一个 page-wise 的推荐框架 DeepPage，该框架通过结合强化学习中的 actor-critic 算法，通过用户的实时反馈来优化商品推荐；Zhao 等[9]提出一种新的强化学习框架 DEERS，通过同时考虑到正反馈和负反馈来优化推荐结果，其中正反馈是指用户点击了系统推荐的商品，负反馈是指用户跳过了系统推荐的商品[3]；Zou 等[52]考虑到现有的推荐系统一般是把提高商品的点击通过率（CTR）作为目标而忽略了用户长期满意度的优化，而长期满意度可以表现在用户浏览的持续时间、用户活跃度等，因此提出了优化用户的延时反馈矩阵，并设计了一个 S 网络来模拟真实环境中用户的反馈。

（2）新闻领域

由于新闻的动态变化性以及推荐的时效性，通过强化学习不仅可以习得短期以及长期阶段内用户的兴趣变化以做出个性化的、针对性的推荐，同时可以达到长期奖励最大化的效果。雅虎实验室曾提出应用 Contextual Bandits 方法于个性化新闻推荐中，采用 Lin UCB 算法达到了较高的计算效率以及较为精准的推荐结果；Shen 等[44]在 Contextual Bandits 的基础上加入一种新的深度神经记忆增强机制来解决使用强化学习方法推荐时需要用户与系统进行大量交互的问题，其为每个用户建模和跟踪历史状态，可以通过少量的交互快速了解用户的兴趣；Zheng 等[37]使用了 DQN（Deep Q network）网络来捕捉用户对新闻的动态兴趣，训练时使用 DDPG，把用户活跃度作为奖励的补充以更好地获取用户偏好，并且加入新的 exploration

策略以避免向用户推荐过多的重复商品。

（3）医疗领域

医疗领域是近年来强化推荐研究的一个重要领域。Zhang 等[58]提出 LEAP 算法，将治疗推荐分解为一个连续的决策过程，并使用循环解码器对标签依赖关系建模，将外部的临床知识融入到奖励的设计中，以有效地防止产生不良的药物组合。Wang 等[65]提出了应用 actor-critic 来处理多种药物、疾病和个人特征之间的复杂关系，指标信号和评估信号两者共同调整其行为的更新方向，以确保有效的处方和低死亡率；然后进一步使用 RNN 来解决由于在现实生活中可能不能观察到全部状态而导致的部分观测马尔可夫过程（Partially Observable Markov Decision Process, POMDP）。

（4）音乐领域

除了电商领域外，音乐也是强化推荐的重点应用领域。Liebman 等[39]采用了 model-based 基于 MDP 的强化学习方法来进行音乐列表的推荐，他们不仅把用户对音乐内容（如旋律、歌词等）的反馈作为奖励，同时通过强化学习方法学习用户歌单列表顺序的偏好；Wang[94]在环境中加入了经加权矩阵分解与卷积神经网络提取歌曲的音频特征，智能体能够学习歌曲的音频转换来挖掘用户偏好。Hong 等[95]利用无线传感器收集用户听歌时的生理信号，如心率和呼吸，来辅助智能体决策，更适用于运动场景。

（5）旅游领域

旅游也是人们生活中涉及到的娱乐项目之一，而目前关于旅游景点推荐的研究较少。Massimo 等[96]把逆强化学习应用在旅游景点推荐中，利用用户过去的景点选择顺序和当时的场景上下文来理解用户的偏好，建立了一个考虑到商品消费时序性的偏好学习模型，然后进一步使用逆强化学习方法进行旅游景点推荐。

基于强化学习的推荐模型分类如表 1 所列。

7 基于强化学习的推荐研究展望

近年来，大数据、人工智能、深度学习技术的出现增强了人机交互，为强化学习在推荐系统中的应用提供了重要的源数据和技术支持。未来，强化学习在推荐系统的研究与应用将主要表现在以下几个方面。

（1）大规模动作空间的强化学习推荐架构

推荐环境对比于其他环境（如围棋只有上、下、左、右 4 个动作）来说动作空间规模巨大，从海量数据中找到用户喜欢的商品需要智能体多次的摸索，导致模型训练很难收敛，收敛稳定性差[62]，进而使得模型大概率学习不到最优策略。同时推荐系统不可能实时获取用户轨迹，实时更新策略，导致策略的更新来自历史用户轨迹，进而使策略更新产生偏差[61]。因此，对于大规模动作空间的推荐系统探索出新的强化学习模型至关重要。此外，基于无模型学习（model-free）的深度强化学习方法通常需要大量的线上交互，通过用户在线反馈训练推荐策略。该过程会消耗大量交互成本，影响用户体验。相对来说，有模型学习（model-based）的方法不需要大量交互[72]，但目前关于这方面的研究刚刚起步，还需要学者们更加深入的研究。

（2）构造推荐环境的模拟仿真器以降低训练成本

由于强化学习方法需要基于现实的交互环境，对于推荐环境来说，其动作空间巨大，线上训练时数据量巨大，对于计算机硬件要求高；并且深度强化学习在线上训练的用户成本很高[52]。因为将一个不成熟的策略在线上训练可能会给用户推荐很多不感兴趣的商品，从而导致用户流失率增多。若使用环境仿真器进行线下训练则会大大降低成本[83]。目前已经有很多研究者使用自己设计的推荐环境仿真器来为训练评估提供支持[52]。也有研究者开发出针对特定数据集的环境模拟器[83][84]，或针对开放数据集（如 movielens 等）的通用环境模拟器[86]。虽然现在已有不少研究成果，但这个方向仍值得更多的学者加以关注。

表 1 基于强化学习的推荐模型分类

Table 1 Classification of reinforcement learning based recommendation models

类别	传统强化学习推荐		深度强化学习推荐	
	基于 MAB 的推荐	**基于 MDP 的推荐**	**基于值函数 DRL 的推荐**	**基于策略梯度 DRL 的推荐**
输入	动作 / 状态和动作	状态和动作 / 状态	状态和动作 / 状态	状态和动作 / 状态
输出	奖励值	Q 值	Q 值	采取某个动作的概率
特征	最大化单步奖励	最大化总奖励	最大化总奖励	最大化总奖励
文献模型	MusicCN-Bandit[39] DCdrift[97] CoLin[40] Corr-Bandit[42] MF-Bandit[43] e-TSbandit [41] DMCB [44] POMDP-Rec[98] Web-Bandit[24]	constrained PSRL[48] Multi_With[45] Bic-RL[47] Bayes-UCB-CN[39] ε-SVR-C[97] DPG-FBE [99] DJ-MC [46] APG[100] IRL-based[96]	DRN[37] Robust DQN[51] DEERS[9] FeedRec[52] GAUM[73] SLATEQ[54] Pseudo Dyna-Q[55] Value-aware[53] UDQN[56] GCQN[57]	LEAP[58] DeepPage[67] CRS[60] DeepChain[72] Top-k Corrected RCR[68] REINFORCE[61] MARDPG[80] LSIC[59] VL-Rec[69] TPGR[62] MaHRL[77] PGCR[63] HRL-Rec[78] IRecGAN[64] KERL[70] SRL-RNN[65] KGRL[71] LIRD[66] NRRS[101]

（3）深度学习技术增强动作状态与奖励的表征

在强化学习中需要研究者自己定义马尔可夫过程中的各个要素，如动作、状态、奖励。状态作为输入特征之一，其构建对于模型训练至关重要。目前的研究对于状态的定义大多使用循环神经网络或者注意力模型、记忆网络等捕捉序列关系[55]。奖励函数作为策略更新的方向，其定义同样是模型训练时的重要一环。而目前对于奖励函数的定义大都是通过手工定义的，对于推荐系统来说统一的奖励定义并不能准确地反映出不同的用户满意度。例如对于围棋或者一些游戏（如 Atari[102] ）来说，其规则固定因此奖励函数也是固定的，而推荐环境中用户的满意度就是奖励函数，用户的满意度表现多种多样。但目前的研究大都是通过直接手工定义奖励（如购买为 5，加入购物车为 3，忽略为 0 等），不能准确地捕捉用户偏好。有学者使用对抗生成网络来模拟奖励函数[73]，实验结果表明相较于手工定义的奖励来说，这种方式更加准确。因此，利用深度学习技术获得状态、奖励等特征的深层表示是未来研究的关注点。

（4）建立更合理的交互性推荐评价机制

目前，学者们对强化学习推荐模型进行评估时使用的评估指标大都是普遍的指标，如 CTR、准确率、召回率、NDCG、MAP 等[103]，这些指标对于基于交互性推荐系统来说并不能准确反映用户对整个交互过程以及推荐结果的满意度。有学者使用 3 个指标来评价推荐算法的好坏[52]：每个会话中平均的点击数（所有用户在一个会话中的点击数的平均值）、每个会话的平均深度（每个用户的会话浏览深度的平均值）、平均回访时间（用户两次会话之间间隔时间的平均值）。这些评估指标不仅考虑了交互过程中用户的短期满意度，还考虑到他们的长期满意度。但这些评估方法并没有针对特定的应用场景做出用户满意度的真实评判，比如对于电商推荐系统来说，用户加入购物车购买等行为也极大地体现了他们的满意度；对于音乐推荐系统来说，用户的听歌时长而不仅仅是是否听了这首歌也体现了用户的效用。因此，在考虑动态推荐环境的情况下，学者们更需要关注如何构造能够反映用户满意度的新评价指标。

（5）强化学习推荐研究呈现应用领域多样化

现有的推荐研究应用领域多种多样[104]，比如社会化推荐、跨领域推荐、组推荐、可解释性推、序列推荐和对话推荐等。而现有的基于强化学习的推荐研究还有待在这些领域深入探究，后续有望将基于强化学习的推荐研究应用于各个细分的推荐领域中。

参考文献

[1] MARZ N, WARREN J. Big Data: Principles and best practices of scalable realtime data systems [M]. USA: Manning, 2015: 44-49.

[2] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems [J]. Computer, 2009, 42(8): 30-37.

[3] BOBADILLA J, ORTEGA F, HERNANDO A, et al. Recommender systems survey [J]. Knowledge Based Systems，2013, 46(Complete):109–132.

[4] HUANG L W, JIANG B T, LV S Y, et al. Survey on deep learning based recommender systems [J]. Chinese Journal of Computers, 2018, 41(7): 1619-1647. (黄立威, 江碧涛, 吕守业, 等. 基于深度学习的推荐系统研究综述 [J]. 计算机学报, 2018, 41(7): 1619-1647.)

[5] BATMAZ Z, YUREKLI A, BILGE A, et al. A review on deep learning for recommender systems: challenges and remedies [J].

Artificial Intelligence Review, 2019, 52(1): 1–37.

[6] ZHAO X Y, XIA L, TANG J L, et al. Deep Reinforcement Learning for Search, Recommendation and Online Advertising: A Survey [J]. ACM SIGWEB Newsletter, 2019 (Spring): 1-15.

[7] LIU Q, ZHAI J W, ZHANG Z Z, et al. A Survey on Deep Reinforcement Learning. Chinese Journal of Computers, 2018, 41(1): 1-27. (刘全，翟建伟，章宗长，等. 深度强化学习综述 [J]. 计算机学报, 2018, 41(1): 1-27.)

[8] ZHAO X X, ZHANG W N, WANG J. Interactive collaborative filtering [C] // Proceedings of the 22nd ACM international conference on information & knowledge management. ACM Press, 2013: 1411–1420.

[9] ZHAO X Y, ZHANG L, DING Z Y, et al. Recommendations with Negative Feedback via Pairwise Deep Reinforcement Learning [C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1040–1048.

[10] WAN L P, LAN X G, ZHANG H B. The theory and application of deep reinforcement learning [J]. Pattern recognition and artificial intelligence, 2019, 32(1):67-81. (万里鹏，兰旭光，张翰博. 深度强化学习理论及其应用综述 [J]. 模式识别与人工智能, 2019, 032(001):67-81).

[11] SARWAR B M, KARYPIS G, KONSTAN J A, et al. Item-based collaborative filtering recommendation algorithms [C] // Proceedings of the 10th international conference on world wide web, 2001: 285–295

[12] VAN M R, VAN S M. Using content-based filtering for recommendation [C] // Proceedings of the Workshop on Machine Learning in The New Information Age, 2000: 47–56

[13] AN M X, WU F Z, WU C H, Neural News Recommendation with Long- and Short-term User Representations [C]// The 57th Annual Meeting of the Association for Computational Linguistics, 2019: 336-345

[14] MA J Q, ZHAO Z, YI X Y. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1930-1939

[15] DU W, DING S F, A survey of Multi-Agent Reinforcement Learning [J]. Computer Science, 2019, 46(8): 1-8. (杜威，丁世飞. 多智能体强化学习综述 [J]. 计算机科学, 2019, 46(8): 1-8.)

[16] LIN X, CHEN H J, A pareto-efficient algorithm for multiple objective optimization in e-commerce recommendation [C] //Proceedings of the 13th ACM Conference on Recommender Systems, 2019: 20-28

[17] GE T Z, ZHAO L Q. Image Matters: Visually modeling user behaviors using Advanced Model Server [C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. 2018: 2087-2095.

[18] GUO Q Y, ZHUANG F Z, QIN C, et al. A survey on knowledge graph-based recommender systems [J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 50(7): 937-957,

[19] YUE Y S, GUESTRIN G. Linear submodular bandits and their application to diversified retrieval [C]// Neural Information Processing Systems. 2011: 2483–2491.

[20] SHANI G, HECKERMAN D, BRAFMAN R I. An MDP-based recommender system [J]. Journal of Machine Learning Research, 2005, 6(9): 1265–1295

[21] AUER P. Using confidence bounds for exploitation-exploration trade-offs [J]. Journal of Machine Learning Research, 2002, 3(1): 397-422.

[22] AGRAWAL S, GOYAL N. Analysis of thompson sampling for the multi-armed bandit problem [C]// Proceedings of the 25th Annual Conference on Learning Theory. 2012: 39.1-39.26

[23] BOUNEFFOUF D, BOUZEGHOUB A, GANCARSKI AL. A contextual-bandit algorithm for mobile context-aware recommender system [C]// Neural Information Processing, 2012: 324–331.

[24] LI L H, CHU W, LANGFORD J, et al. A Contextual-Bandit Approach to Personalized News Article Recommendation [C] // Proceedings of the 19th international conference on World wide web. Raleigh, 2010: 661-670

[25] ALLESIARDO R, FERAUD R, BOUNEFFOUF D. A neural networks committee for the contextual bandit problem [C]// International Conference on Neural Information Processing. 2014: 374-381.

[26] AGRAWAL S, GOYAL N. Thompson sampling for contextual bandits with linear payoff s [C]// International Conference on Machine Learning. 2013: 127-135.

[27] LIU J W, GAOF, LUO X L, A survey of deep reinforcement learning based on value function and strategy gradient [J]. Journal of Computer Science, 2019, 42(6): 1406-1438. (刘建伟, 高峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述 [J]. 计算机学报, 2019, 42(6): 1406-1438.)

[28] MNIHL V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning [J]. Nature, 2015, 518(7540): 529-542.

[29] VAN H H, GUEZAND A, SILVER D. Deep Reinforcement Learning with Double Q-learning [C]// Proceedings of AAAI conference on Artificial Intelligence. 2016: 2094-2110.

[30] WANG Z Y, SCHAUL T, HESSEL M, et al. Dueling Network Architectures for Deep Reinforcement Learning [C]// International conference on machine learning. 2016: 1995-2003.

[31] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized Experience Replay [C]// Proceedings of international conference on learning representations, 2016: 1-21.

[32] FORTUNATO M, AZARM G, PIOT B, et al. Noisy networks for exploration [J]. arXiv:1706.10295, 2017.

[33] BELLEMARE M G, DABNEY W, MUNOS R. A distributional perspective on reinforcement learning [C]// International Conference on Machine Learning. 2017: 449-458.

[34] HESSEL M , MODAYIL J , VAN H H , et al. Rainbow: Combining Improvements in Deep Reinforcement Learning [C]// Proceedings of Association for the Advancement of Artificial Intelligence. 2018: 3215-3222.

[35] SILVER D, LEVER G, HEESS N, et al. Deterministic Policy Gradient Algorithms [C]// International conference on machine learning. 2014: 387-395.

[36] KULKARNI T D, NARASIMHAN K R, SAEEDI A, et al. Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation [C]// Proceedings of Thirtieth Conference on Neural Information Processing Systems, 2016: 1-9

[37] ZHENG G J, ZHANG F Z, ZHENG Z H, et al. DRN: A Deep Reinforcement Learning Framework for News Recommendation [C]// Proceedings of the 2018 World Wide Web Conference, 2018: 167–176.

[38] SHANI G, GUNAWARDANA A. Evaluating recommendation systems [M]. Recommender systems handbook. Boston: Springer, 2011: 257–297.

[39] WANG X X, WANG Y, HSU D, et al. Exploration in Interactive Personalized Music Recommendation: A Reinforcement Learning Approach [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2014, 11(1): 1-22

[40] WU Q Y, WANG H Z, GU Q Q, et al. Contextual Bandits in a Collaborative Environment [C]// Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, 2016: 529–538

[41] BRODEN B, HAMMAR M, NILSSON B J, et al. Ensemble Recommendations via Thompson Sampling: an Experimental Study within e-Commerce [C]// Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, 2018: 19–29.

[42] WANG H Z, WU Q Y, WANG H N. Factorization Bandits for Interactive Recommendation [C]// Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017: 2695-2702.

[43] INTAYOAD W, KAMYOD C, TEMDEE P. Reinforcement Learning Based on Contextual Bandits for Personalized Online Learning Recommendation Systems [J]. Wireless Personal Communications, 2020(115): 2917-2932.

[44] SHEN Y L, DENG Y, RAY A, et al. Interactive recommendation via deep neural memory augmented contextual bandits [C]// Proceedings of the 12th ACM conference on recommender systems, 2018: 122–130.

[45] ZHANG Y, ZHANG C W, LIU X Z. Dynamic Scholarly Collaborator Recommendation via Competitive Multi-Agent Reinforcement Learning [C]// Proceedings of the Eleventh ACM Conference on Recommender Systems, 2017: 331–335.

[46] LIEBMAN E, SAAR T M, STONE P. DJ-MC: A Reinforcement-Learning Agent for Music Playlist Recommendation [C]// Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems, 2015: 591-599.

[47] SUNGWOON C, HEONSEOK H, UIWON H, et al. Reinforcement Learning based Recommender System using Biclustering Technique [J]. arXiv:1801.05532, 2018.

[48] DE N F, THEOCHAROUS G, VLASSIS N, et al. Capacity-aware Sequential Recommendations [C]// Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems, 2018: 416-424.

[49] LIU W, LIUF, TANG R, et al. Balancing Between Accuracy and Fairness for Interactive Recommendation with Reinforcement Learning[C] // Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2020: 155-167.

[50] LIU F, TANG R, GUO H, et al. Top-aware reinforcement learning based recommendation [J]. Neurocomputing, 2020, 417: 255-269.

[51] CHEN S Y, YU Y, DA Q, et al. Stabilizing Reinforcement Learning in Dynamic Environment with Application to Online Recommendation [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 1187–1196.

[52] ZOU L X, XIA L, DING Z Y, et al. Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 2810-2818

[53] CHANG H P, YANG X R, CUI Q, et al. Value-aware Recommendation based on Reinforcement Profit Maximization [C]// Proceedings of the 2019 World Wide Web Conference, 2019: 3123-3129.

[54] IE E, JAIN V, WANG J, et al. Slate Q: a tractable decomposition for reinforcement learning with recommendation sets [C]// Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019: 2592-2599.

[55] FENG S X, CHEN H S, LI K, et al. Pseudo dyna-q: a reinforcement learning framework for interactive recommendation [C]// Proceedings of the 13th International Conference on Web Search and Data Mining, 2020: 816–824.

[56] LEI Y, LI W J. Interactive Recommendation with User-Specific Deep Reinforcement Learning [J]. ACM Transactions on Knowledge Discovery from Data, 2019, 13(6): 1-15.

[57] LEI Y, PEI H, YAN H, et al. Reinforcement learning based recommendation with graph convolutional q-network [C]// Proceedings

of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020: 1757-1760.

[58] ZHANG Y T, CHEN R, TANG J, et al, LEAP: Learning to Prescribe Effective and Safe Treatment Combinations for Multimorbidity [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017: 1315–1324.

[59] ZHAO W, WANG W Y, YE J B, et al. Leveraging Long and Short-Term Information in Content-Aware Movie Recommendation via Adversarial Training [J]. IEEE Transactions on Cybernetics, 2019, 50(11): 4680-4693.

[60] SUN Y M, ZHANG Y. Conversational Recommender System [C]// Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018: 235–244

[61] CHEN M M, BEUTEL A, COVINGTON P, et al. Top-K Off-Policy Correction for a REINFORCE Recommender System [C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019: 456-464.

[62] CHEN H K, DAI X Y, CAI H, et al. Large-scale interactive recommendation with tree-structured policy gradient [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 3312-3320.

[63] PAN F Y, CAI Q P, TANG P Z, et al. Policy gradients for contextual recommendations [C]// Proceedings of The World Wide Web Conference, 2019: 1421–1431.

[64] BAI X Y, GUAN J, WANG H N. A model-based reinforcement learning with adversarial training for online recommendation [C]// Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019: 1-12.

[65] WANG L, ZHANG W, HE X F, Supervised Reinforcement Learning with Recurrent Neural Network for Dynamic Treatment Recommendation [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 2447-2456

[66] ZHAO X Y, XIA L, ZHANG L, et al. Deep Reinforcement Learning for List-wise Recommendations [J]. arXiv:1801.00209, 2017.

[67] ZHAO X Y, XIA L, ZHANG L, et al. Deep Reinforcement Learning for Page-wise Recommendations [C]// Proceedings of the 12th ACM Conference on Recommender Systems, 2018: 95–103.

[68] ZHANG R Y, YU T, SHEN Y L, et al. Text-based interactive recommendation via constraint-augmented reinforcement learning [C]// Proceedings of the 33rd Conference on Neural Information Processing Systems, 2019: 13-24.

[69] YU T, SHEN Y L, ZHANG R Y, et al. Vision-Language Recommendation via Attribute Augmented Multimodal Reinforcement Learning [C]// Proceedings of the 27th ACM International Conference on Multimedia, 2019: 39–47.

[70] WANG P, FAN Y, XIA L, et al. KERL: A knowledge-guided reinforcement learning model for sequential recommendation [C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 209-218.

[71] CHEN X, HUANG C, YAO L, et al. Knowledge-guided deep reinforcement learning for interactive recommendation [C]// Proceedings of the 2020 International Joint Conference on Neural Networks, 2020: 1-8.

[72] ZHAO X Y, XIA L, ZHANG L, et al. Model-Based Reinforcement Learning for Whole-Chain Recommendations [J]. arXiv:1902.03987, 2019.

[73] CHEN X S, LI S, LI H, et al. Generative Adversarial User Model for Reinforcement Learning Based Recommendation System [C]// Proceedings of the 34th International Conference on Machine Learning, 2019: 1052-1061.

[74] GAO R, XIA H F, LI J, et al. DRCGR: Deep Reinforcement Learning Framework Incorporating CNN and GAN-Based for Interactive Recommendation [C]// Proceedings of the 2019 IEEE International

Conference on Data Mining, 2019: 1048-1053.

[75] WU H J, DAI D D, FU Q M. Research Progress on the combination of reinforcement learning and generative adversary network [J]. Journal of Computer engineering and Application, 2019, 55(10):41-49. (吴宏杰，戴大东，傅启明. 强化学习与生成式对抗网络结合方法研究进展 [J]. 计算机工程与应用, 2019, 55(10):41-49)

[76] LIN J H, ZHANG Z C, JIANG C. A survey of imitation learning based on generative adversary network [J], Journal of Computer Science, 2020,43(2): 326-351. (林嘉豪，章宗长，姜冲. 基于生成对抗网络的模仿学习综述 [J]. 计算机学报, 2020, 43(2): 326-351.)

[77] ZHAO D Y, ZHANG L, ZHANG B, et al. MaHRL: Multi-goals Abstraction Based Deep Hierarchical Reinforcement Learning for Recommendations [C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020: 871-880.

[78] XIE R B, ZHANG S L, WANG R, et al. Hierarchical Reinforcement Learning for Integrated Recommendation [C]// Proceedings of the 35th AAAI Conference on Artificial Intelligence, 2021: 1-8.

[79] ZHANG J, HAO B W, CHEN B, et al. Hierarchical reinforcement learning for course recommendation in MOOCs [C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 435–442.

[80] FENG J, LI H, HUANG M, et al. Learning to collaborate: Multi-scenario ranking via multi-agent reinforcement learning [C]// Proceedings of the World Wide Web Conference, 2018: 1939-1948.

[81] HE X, AN B, LI Y, et al. Learning to Collaborate in Multi-Module Recommendation via Multi-Agent Reinforcement Learning without Communication[C] // Proceedings of the Fourteenth ACM Conference on Recommender Systems. 2020: 210-219

[82] GUI T, LIU P, ZHANG Q, et al. Mention Recommendation in Twitter with Cooperative Multi-Agent Reinforcement Learning [C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019: 535-544

[83] SHI J C, YU Y, DA Q, et al. Virtual-Taobao: virtualizing real-world online retail environment for reinforcement learning [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33(01): 4902-4909.

[84] SHANG W J, YU Y, LI Q Y, et al. Environment reconstruction with hidden confounders for reinforcement learning based recommendation [C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019: 566-576.

[85] ZHAO X Y, XIA L, ZOU L X, et al. Toward simulating environments in reinforcement learning based recommendations [J]. arXiv:1906.11462, 2019.

[86] ROHDE D, BONNER S, DUNLOP T, et al. RecoGym: a reinforcement learning environment for the problem of product recommendation in online advertising [J]. arXiv:1808.00720, 2018.

[87] SHI B, OZSOY M G, HURLEY N, et al. PyRecGym: a reinforcement learning gym for recommender systems [C]// Proceedings of the 13th ACM Conference on Recommender Systems, 2019: 491-495.

[88] IE E, HSU C, MLADENOV M, et al. RecSim: a configurable simulation platform for recommender systems [J]. arXiv:1909.04847, 2019.

[89] WANG X T, CHEN Y R, JIE Y, et al. A reinforcement learning framework for explainable recommendation [C]// Proceedings of the 2018 IEEE International Conference on Data Mining, 2018: 587-596.

[90] XIAN Y K, FU Z H, MUTHUKRISHNAN S. Reinforcement knowledge graph reasoning for explainable recommendation [C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019: 285–294.

[91] MCINERNEY J, LACKER B, HANSEN S, et al. Explore, exploit, and explain: personalizing explainable recommendations with bandits [C]// Proceedings of the 12th ACM Conference on Recommender Systems, 2018: 31–39.

[92] LEI Y, WANG Z T, LI W J. Social attentive deep q-network for recommendation [C]// Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019: 1189–1192.

[93] LIU F, GUO H F, LI X T, et al. End-to-end deep reinforcement learning based recommendation with supervised embedding [C]// Proceedings of the 13th International Conference on Web Search and Data Mining, 2020: 384–392.

[94] WANG Y. A Hybrid Recommendation for Music Based on Reinforcement Learning [C]// Pacific-Asia Conference on Knowledge Discovery and Data, 2020: 91-103.

[95] HONG D, LI Y, DONG Q. Nonintrusive-Sensing and Reinforcement-Learning Based Adaptive Personalized Music Recommendation [C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020: 1721-1724.

[96] MASSIMO D, ELAHI M, RICCI F. Learning User Preferences by Observing User-Items Interactions in an IoT Augmented Space [C]// Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization, 2017: 35–40.

[97] ZHAO Y, ZENG D, SOCINSKI M A, et al.. Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer [J]. Journal of the International biometric society, 2011, 67(4): 1422–1433.

[98] LU Z Q, YANG Q, Partially Observable Markov Decision Process for Recommender Systems [J]. arXiv:1608.07793, 2016.

[99] HU Y J, DA Q, ZENG A X, et al. Reinforcement Learning to Rank in E-Commerce Search Engine: Formalization, Analysis, and Application [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018: 368-377.

[100] CHI C Y, TSAI R T H, LAI J Y, et al. A Reinforcement Learning Approach to Emotion-based Automatic Playlist Generation [C]// Proceedings of International Conference on Technologies and Applications of Artificial Intelligence, 2010: 60-65.

[101] ZENG C Q, WANG Q, MOKHTARI S. Online Context-Aware Recommendation with Time Varying Multi-Armed Bandit [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 2025–203.

[102] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with Deep Reinforcement Learning [C]// Proceedings of the Workshops at the 26th Neural Information Processing Systems, 2013: 201-220.

[103] ZHU Y X, LV L Y. Evaluation Metrics for Recommender Systems [J]. Journal of University of Electronic Science and Technology of China, 2012, 41(2): 163-176.

[104] ZHANG S, YAO L N, SUN A X, et al. Deep learning based recommender system: a survey and new perspectives [J]. ACM Computing Surveys, 2019, 52(1): 1–38.

作者简介

Yu Li, born in 1973, PH.D., associate professor. His main research interests include deep learning, recommender systems.

余力 (通讯作者)，博士，副教授，主要研究领域为推荐系统、深度学习.

杜启翰，博士研究生，主要研究领域为推荐系统、深度学习.

岳博妍，硕士研究生，主要研究领域为推荐系统、强化学习.

向君瑶，硕士研究生，主要研究领域为推荐系统、深度学习.

徐冠宇，本科生，主要研究领域为深度学习.

冷友方，博士研究生，主要研究领域为推荐系统，深度学习.