文章编号:1671-4598(2025)08-0266-08 DOI:10.16526/j.cnki.11-4762/tp.2025.08.033 中图分类号:TP391 文献标识码:A

基于强化学习的知识图谱产品推荐方法

王 快[1,2],张 骞[3],易元刚[1]
(1.重庆市测绘科学技术研究院,重庆 401120;
2.自然资源部智能城市时空信息与装备工程技术创新中心,重庆 401120;
3.深圳大学 建筑与城市规划学院,广东 深圳 518060)

摘要:考虑产品环境感知状态选择对后续行为的影响,提出基于强化学习的知识图谱推荐方法;该方法构建一个强化学习知识推荐框架,结合短期奖励与长期增量评价提出双重奖励驱动机制,用于改进演员-评论家网络的全局推理能力;利用增强损失约束作为信号监督策略梯度更新,形成自监督路径推理策略;通过实验证明:作为改进实体知识推荐实例,所提方法相比基准学习方法具有较高的推荐精度和平均奖励,能够引导策略寻找推荐路径并提供有效解释,为知识推理提供了决策支持。

关键词:强化学习;知识图谱;双重奖励驱动;自监督推理;增强损失约束

Knowledge Graph Product Recommendation Method Based on Reinforcement Learning

WANG Kuai[1,2], ZHANG Qian[3], YI Yuangang[1]
(1. Chongqing Academy of Surveying and Mapping, Chongqing 401120, China;
2. Technology Innovation Center for Spatio-temporal Information and Equipment of Intelligent City, Minisity pf Natural Resources, Chongqing 401120, China;
3. College of Architecture and Urban Planning, Shenzhen University, Shenzhen 518060, China)

Abstract: To consider the impact of current environmental perception state selection on subsequent behaviors, a knowledge graph recommendation method based on reinforcement learning is proposed. This method constructs a reinforcement learning framework, and presents a dual reward driven strategy, which combines short-term rewards with long-term incremental evaluations to improve the global reasoning ability of the actor-critic network; The method utilizes the enhanced loss constraints as signal supervision strategy for gradient updates, achieving a self supervised path inference strategy. Experimental results show that, as an example of improving knowledge recommendation, the proposed model is superior to the baseline methods in the recommendation accuracy and average reward, and can guide strategies to find out recommendation paths and provide effective explanations, providing decision support for path reasoning.

Keywords: reinforcement learning; knowledge graph; dual reward driven; self supervised reasoning; enhanced loss constraints

0 引言

面向知识图谱的产品推荐方法[1]通过从大规模数据中学习实体之间的关系,旨在提高推荐系统的可信度,受到国内外学者广泛关注[2-3]。其方法[4]可分为嵌入式、路径搜索和学习推理。其中,嵌入式方法通过实体距离表示知识图谱中节点的关系,包括贝叶斯个性化排序[5](BPR,Bayesian personalized ranking)、可迁移推荐[6](TransRec,transferable recommendation)等方法;路径搜索方法通过元路径在知识图谱上进行推理,包括协同过滤[7](CF,collaborative filtering)、策略引导路径推理[8](PGPR,policy-guide path reasoning)等方法;

收稿日期:2024-07-02; **修回日期**:2024-08-14。
基金项目:国家自然科学基金(41801274);重庆市科研机构绩效激励引导专项项目(STB2023JXJL-YFX0048)。
作者简介:王 快(1987-),女,硕士,工程师。
通讯作者:张 骞(1987-),男,硕士,副教授。
引用格式:王 快,张 骞,易元刚.基于强化学习的知识图谱产品推荐方法[J].计算机测量与控制,2025,33(8):266-273,282.

学习推理方法利用学习算法生成知识网络进而对实体关系进行推理和应用，包括联合表示学习[9]（JRL，joint representation learning）、深度协作神经网络[10]（DeepConn，deep cooperative neural network）等方法。然而，在处理产品知识图谱中的多源实体关系时，这些方法由于缺少辅助模型生成有效的搜索策略，较难寻找不连续实体之间的关系[11]，在执行显式推理时容易陷入诸如局部最优等困扰[12-13]。

强化学习[14]（RL，reinforcement learning）是一种从试错过程中发现最优行为策略的技术，利用马尔科夫决策过程将知识图谱推理任务转化为序贯任务决策过程，通过训练 Agent[15]将产品源实体连接到目标实体的一系列行为，实现产品知识的可解释性推荐。Agent 作为实体之间环境交互的可解释性接口[16]，引导实体发现推荐路径。文献［17］将强化学习引入产品知识图谱中，增强知识的状态表示，设计了一个复合奖励函数计算序列和知识级别的奖励；文献［18］将强化学习与注意力保持机制相结合，形成多跳推理模型，以解决时序知识图谱缺乏路径记忆组件的问题；文献［19］利用专家路径演示的对抗性，提出基于演员－评论家网络的强化学习知识推荐模型，以解决不连续实体状态下的产品实体连接学习问题。从问题解决的角度来看，以上文献更多考虑当前状态下的奖励，忽略了后续状态下根据时间差异增量选择行为的影响。因此，部分研究者更加关注强化学习的行为奖励[20]，文献［21］遵循多跳评分函数和 SoftMax 机制，提出了策略引导的路径推理方法，用于采样个性化推荐路径，但该方法缺乏辅助策略搜索潜在路径，导致生成可解释路径时遗漏了有效路径。同时，大部分现有强化学习方法在搜索较短的实体关系路径时依据最终或更多的奖励更新策略，忽略了知识图谱的网络评估能力[22]，因此如果推荐策略对状态的估值高于真实值，模型将不能有效地学习路径搜索策略，从而降低推荐精度。

基于上述问题描述，本文提出一种基于强化学习的知识图谱产品推荐方法（KGR _ RL，knowledge graph product recommendation based on reinforcement learning），旨在引导策略提供可解释性推荐路径，为知识推理提供决策支持。

1　问题表示与框架设计

1.1　问题表示

面向知识图谱的产品推荐将具有关系集 R 的知识图谱 G_R 定义为：

$$G_R=\{(e_{\text{head}},r,e_{\text{tail}})\mid e_{\text{head}},e_{\text{tail}}\in E,r\in R\}\quad(1)$$

式中，e_{head}为头部实体，e_{tail}为尾部实体，三元组（e_{head}，r，e_{tail}）为通过关系 $r\in R$ 建立从 e_{head}到 e_{tail}的知识图谱 G_R，G_R 包含实体集合 E，E 包含一系列用户实体 U 和项目实体 I，通过生成多跳路径连接在知识图谱中建立关系 $r_{u,i}\in R$，u 为推荐系统中的用户，i 为被推荐至用户的产品，实体路径推理的元路径对应知识解释策略。

1.2　框架设计

根据知识图谱 G_R，构建了一个强化学习知识推荐框架。如图 1 所示，该框架由环境感知、双重奖励驱动机制和自监督强化学习 3 个部分组成。其中，环境感知通过马尔可夫决策过程对知识图谱的实体和关系进行建模，包括状态、动作、状态转移概率、奖励和折现系数；双重奖励驱动机制结合短期奖励与长期增量评价，用于改进演员－评论家网络，演员网络用于学习动作空间 A_t 和状态空间 S_t 上的推荐路径策略 π_φ，评论家网络用于评估状态值并生成状态值函数 $\tilde{v}(S_t)$ 的估计值；自监督强化学习利用增强损失约束作为信号监督策略梯度更新，用于执行知识图谱的实体路径推理。

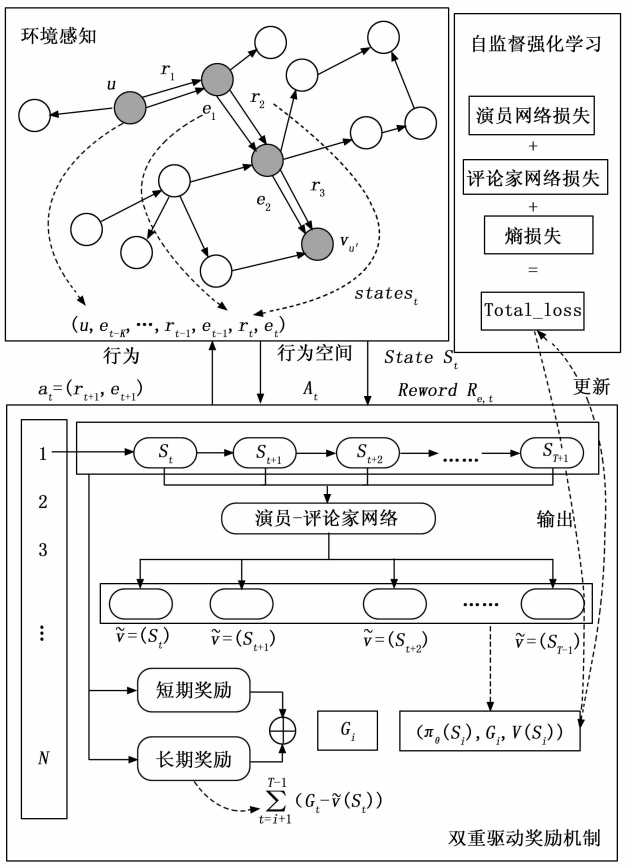


图 1　基于强化学习的知识图谱产品推荐框架

强化学习知识推荐框架，使用短期奖励和长期增量评估来推断未来的路径。首先，运用马尔可夫决策过程，将当前状态 S_t 馈送到演员－评论家网络，以生成当前返回状态 G_t 和评估值 $v(S_t)$，形成具有元路径的初始演员网络生成有限推荐路径；然后，计算当前状态下的时间差增量［$G_t-v(S_t)$］，并将所有后续状态下

的时间差增量累积为长期增量评估，形成以短期奖励与长期增量评估相结合的双重奖励驱动机制；最后，引入增强损失约束自监督环境感知的推理，形成有效的知识推荐。

2 环境感知

合理感知环境状态可以帮助强化学习方法更好地理解知识的状态和采取的动作。因此，产品知识推荐的时序状态需要由离散时间随机决策机制予以支持，马尔可夫决策过程（MDP，Markov decision process）作为序贯决策的过程，能够智能感知当前产品知识的环境状态，利用 Agent 将产品实体与当前环境进行信息交互，包含状态、动作、状态转移概率、奖励和折扣因子等 5 个基本要素。

1）状态。状态空间表示产品搜索推荐解决方案的结构和内容，解决方案的结构形式由产品实体的知识单元定义，表示概念元素（例如：栏目、特征、品牌、目录等）及其相互关系（例如：Purchase、Mention、Belong _ to、Produce _ by、Also _ bought 等）。当用户搜索某一产品时，在时间 t 内产品状态 s_t 定义为三元组 (u, e_t, h_t)。其中，u 为在 U 中的用户，e_t 表示在时间 t 后搜索到达的产品实体 Agent，h_t 为时间 t 前的历史信息。将 n 步历史定义为包含在历史 n 步中产品实体和关系的路径序列，即 $\{e_{t-n}, r_{t-n+1}, \cdots, e_{t-1}, r_t\}$。基于用户 u 初始化状态定义为：

$$S_0 = (u, \varphi) \tag{2}$$

式中，φ 为初始化产品状态下的空集，u 为当前用户。

2）行为。行为空间使用存储的案例或规则知识更新加工过程的结构和内容。对于当前产品实体 e_t 的状态 s_t，产品实体 e_t 的 Agent 执行行为 a_t 以到达下一个产品实体 $e_{t+1} * a_t = (e_t, r_{t+1}) \in A_t$，$r_{t+1}$ 连接为 e_t 和 e_{t+1} 之间的关系，e_t 的一组可能行为为其行为空间 A_t，定义为：

$$A_t = \{(r,e) \mid (e_t, r, e) \in G_R, e \notin [e_0, \cdots, e_{t-1}]\} \tag{3}$$

3）状态转移概率。给定当前产品状态 $s_t = (u, e_t, h_t)$，在执行动作 $a_t = (r_{t+1}, e_{t+1})$ 后 Agent 到达下一个状态的转移概率为：

$$P[S_{t+1} = (u, e_{t+1}, h_{t+1}) \mid S_t = (u, e_t, h_t), a_t = (r_{t+1}, e_{t+1})] \tag{4}$$

4）奖励。状态和动行为空间描述了产品推荐解决方案所需的环境，为使解决方案至知识图谱中的案例和规则知识，需建立一个与产品推荐解决方案相匹配的奖励机制，以获取最优策略。在策略引导路径推理中，给定多跳评分函数 $f(u, i)$，使用 SoftMax[21] 机制计算终端奖励 R_T：

$$R_T = \begin{cases} max\left[0, \dfrac{f(u, e_T)}{max_{i \in I} f(u, i)}\right] & \text{if} \quad e_T \in I \\ 0 & \text{否则} \end{cases} \tag{5}$$

式中，R_T 归一化为 [0，1] 范围，短期或单一奖励可能陷入局部最优。为解决此问题，使用了一个由各种未来状态中的时间差增量积分产生的长期增量评估。为获得更多奖励，产品实体的 Agent 考虑当前奖励和未来奖励，将总回报 G_t 定义为：

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T \tag{6}$$

式中，T 为终止状态，$\gamma \in [0, 1]$ 为折现因子，是当前奖励和未来奖励的折现值叠加。

3 知识推荐

通过表示产品实体知识，将产品实体从需求设计到解决方案的转化形成邻接矩阵计算过程。引入双重奖励驱动机制和自监督强化学习，逐步生成符合预期目标并满足现有知识图谱推荐的解决方案。技术人员不断制定出适合指导实际产品推荐的新解决方案，经过验证的解决方案可以直接存储于知识图谱以实现知识积累。

3.1 演员－评论家网络

将搜索环境感知的信息表示为马尔科夫决策过程后，需利用训练演员－评论家网络进行策略更新，以寻求最佳的产品路径推理策略。

1）演员网络。用于学习产品推荐路径推理策略 π_θ，计算当前状态 S_t（π_θ）中每个所选行为 a_t 的概率分布。演员网络输入行为空间和当前产品实体节点的状态，输出行为空间中每个行为状态的概率分布。首先使用掩码操作删除无效行为，然后将结果输入到 SoftMax 奖励机制[21] 中以生成最终行为概率分布。

2）评论家网络。计算对评论家网络在状态 S_t 中的值，在评论家网络中，输入为当前产品知识单元的状态 S_t，利用奖励 R_T 输出状态的值评估 $\hat{v}(S_t)$。

3.2 双重奖励驱动机制

由于演员－评论家网络模型的策略更新只考虑时间差增量的当前状态，短期奖励 G_t 是在每个未来状态中没有时间差异增量反馈的奖励，仅用于 G_t 训练的演员－评论家网络，忽略了当前状态选择对后续状态和相应时间差异增量的影响，容易使产品知识推荐陷入局部最优状态，使知识实体 Agent 在路径搜索开始时选择次优行为。

为改进演员－评论家网络的全局路径推理能力，提出一种双重奖励驱动机制。如图 2 所示，将短期奖励 G_t 与长期增量评估 TD_{LT} 相结合，TD_{LT} 考虑了在策略 π_θ 下对每个未来状态的时间差增量反馈。将双重奖励驱动机制定义为：

$$Dual\_R = G_t + TD_{LT} \tag{9}$$

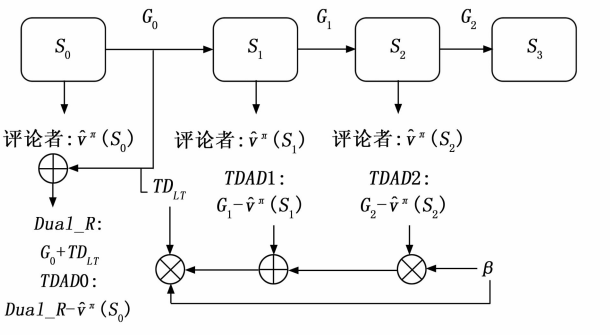


图 2　双重奖励驱动机制

$$TD_{LT}=\sum_{t=t+1}^{T-1}(G_t-\hat{v}(S_t)) \tag{10}$$

式中，$Dual\_R$ 为短期奖励 G_t 与长期增量评估 TD_{LT} 集合，$G_t-\hat{v}(S_t)$ 为对未来状态的时间差增量反馈，将对象的策略更新函数定义为：

$$\nabla_{\Theta}J(\Theta)=E\{\nabla_{\Theta}\log\pi_{\Theta}(S)[Dual\_R-\tilde{v}(S)]\} \tag{11}$$

式中，$Dual\_R-\hat{v}$ (S) 为相应状态的优势函数值，用于表示实际奖励值与状态获得的期望奖励值之间的差距。如果未来状态下对 $Dual\_R$ 值累积衰减大于 $\hat{v}(S_t)$，则说明当前状态对应的行为选择策略质量较高，可以维持；相反，则需要更新策略。因此，双重奖励驱动机制能够有效反映根据当前用户对环境感知的行为选择对未来状态趋势的影响。在知识推荐和路径推理中，双重奖励驱动机制能够确保在选择行为后，当前状态和未来状态的时间差异增量为正，从而在知识图谱中向用户推荐最有潜力的实体项目，并生成更有解释性的反馈路径。

3.3　自监督强化学习知识推荐

为确保强化学习训练策略的评估能力，构建了演员网络、评论家网络的增强损失函数，从而提高知识图谱产品推荐的自监督能力。对于评论家网络，输入状态为 S_t，输出状态为相应的期望值为 $\tilde{v}(S_t)$，损失函数定义为：

$$L_{\text{评论家}}=\sum_{t=0}^{T-1}[Dual_R-\hat{v}(S_t)]^2 \tag{12}$$

对于演员网络，给定状态 S_t 生成模型策略 $\pi_{\theta}(S_t)$，选择对应行为的状态转移概率，为自适应更新执行器网络的参数，将增强损失 L_r 作为约束加入到网络中。损失函数定义如下：

$$L_r=\frac{\sum_{i=0}^{\text{batch\_size}-1}e-\log[\pi_{\theta}(S_t)]*[G_t-\hat{v}(S_t)](i)-1}{batch\_size} \tag{13}$$

$$L_{\text{演员}}=-\log[\pi_{\theta}(S_t)]*[Dual\_R-\hat{v}(S_t)]+L_r \tag{14}$$

式中，$L_{\text{演员}}$ 由动作目标函数和 L_r 两部分组成。其中，动作目标函数基于双重奖励驱动机制，用于获得当前行为选择产生的最大值期望，并利用强化损失 L_r 计算网络的价值评估与行为产生的价值之间的差异。当演员－评论家的价值评估低于行为的实际价值时，则行为选择会带来比预期更多的价值，L_r 会驱动策略选择这类行动，反之则会减少对这类行动的选择。因此，L_r 可以动态调整演员－评论家网络的行为选择策略。在训练强化学习网络时，加入熵损失函数 H (π)，使 Agent 尝试选择一些新的动作。将损失函数定义为：

$$L_{\text{entropy}}=\max_{\pi_{\theta}}H(\pi) \tag{15}$$

然后，增强损失函数计算为：

$$L=L_{\text{演员}}+L_{\text{评论家}}+\beta*L_{\text{entropy}} \tag{16}$$

式中，β 为控制熵损失相对贡献的参数，更新策略梯度 $\nabla_{\theta}J$ (θ) 定义为：

$$\nabla_{\theta}J(\theta)=E_{\pi}\{\nabla_{\theta}\log\pi_{\theta}(S_t)[Dual\_R-\hat{v}(S_t)]\} \tag{17}$$

3.4　训练算法

使用双重奖励驱动机制可以加强对知识图谱的全局推理能力，增强损失约束作为自监督信号可以加强模型训练。首先，初始化演员－评论家网络的参数，在产品知识图谱中预定义元路径和状态三元组。然后，在预定义的元路径下生成具有情节长度 T 的知识推荐路径，将时间 t 内状态 S_t 作为演员和评论家网络的输入，演员网络的输出为评估 $\hat{v}(S_t)$。计算累积产品实体知识推荐奖励 G_t，以及计算从每个状态到结束状态以及每个状态的相应时间差异增量 ($G_t-\hat{v}(S_t)$)。基于未来状态与当前状态时间差异的累积增量长期评估，并结合短期奖励 G_t 来更新策略。然后，更新损失函数，包括评论家损失、演员损失和熵损失。最后，更新策略梯度 $\nabla_{\theta}J$ (θ) 用于更新现有产品知识图谱。算法 KGR_RL 如下所示。

算法：KGR_RL 模型训练

Input：演员网络：$\pi_{\theta}(s)$，评论家网络：$\hat{v}(S)$，迭代次数：N，批大小：K

Output：π_{θ}

Initialize 行为者网络 参数

$n\leftarrow 0$

for n to ($T-1$) do

产生迭代次数 $\pi_{\theta}(s)$：S_0，A_0，R_1，A_1，R_2，…，S_{T-1}，A_{T-1}，R_T

$G_t\leftarrow 0$

for $t\leftarrow 0$ to ($T-1$) do

for $i\leftarrow$ ($T-1$) to t do

$G_t\leftarrow\gamma G_{t+}R_t$

$TD_t=G_t-\hat{v}(S_t)$ end for

$Dual\_R=\sum_{t+1}^{T-1}TD_t+G_t$

$\hat{v}(S_t)\leftarrow\hat{v}(S_t)+\alpha[Dual\_R-\hat{v}(S_t)]$

$L_{\text{评论者}}=\sum_{t=0}^{T-1}(Dual\_R-\hat{v}(S_t))$

$$L_r = \frac{\sum_{i=0}^{\text{batch\_size}-1} e - \log[\pi_\theta(S_t)] * [G_t - \hat{v}(S_t)][i] - 1}{\text{batch\_size}}$$

$$L_{\text{行为者}} = -\log[\pi_\theta(S_t)] * [Dual\_R - \hat{v}(S_t)] + L_r$$

$$L_{\text{entropy}} = \max_{\pi_\theta} H(\pi)$$

$$L_{\text{total}} = L_{\text{演员}} + L_{\text{评论家}} + \beta * L_{\text{entropy}}$$

end for

$$\nabla_\theta J(\theta) = E_\pi[\nabla_\theta \log \pi_\theta(S_t)(Dual\_R - \hat{v}(S_t))]$$

end for

return π_θ

4 实验结果与分析

4.1 实验设置

从中国勘测联合网中获取“海洋测绘设备”和“工程测量设备”两个真实交易集作为实验数据，它们由产品评论、元信息和链接组成，每个数据集包含 8 种关系类型和 5 种实体类型。数据集如表 1 所示，数据类型描述如表 2 所示。

表 1 数据集统计

数据集	用户数	栏目数	关系	实体类型	每个用户购买产品
海洋测绘设备	22 363	12 101	8	5	8.68
工程测量设备	39 387	24 066	8	5	7.16

表 2 数据类型描述

实体	描述
用户	推荐系统中的用户
栏目	被推荐至用户的产品栏目
特征	用户浏览的产品特征
品牌	产品品牌
目录	产品目录
关系	描述
Purchase	User $\xrightarrow{\text{purchase}}$ Item
Mention	User $\xrightarrow{\text{Mention}}$ Feature
Described_by	User $\xrightarrow{\text{Described\_by}}$ Feature
Belong_to	User $\xrightarrow{\text{Belong\_to}}$ Category
Produce_by	User $\xrightarrow{\text{Produce\_by}}$ Brand
Also_bought	User $\xrightarrow{\text{Also\_bought}}$ Item
Also_viewed	User $\xrightarrow{\text{Also\_viewed}}$ another Item
Bought_together	User $\xrightarrow{\text{Bought\_together}}$ another Item

实验包括表征学习和强化学习两个阶段。表征学习中，运用 TransE[23] 模型学习“海洋测绘设备”和“工程测量设备”数据集中的关系和实体嵌入，应用于下游任务中强化学习的训练和奖励计算，将嵌入训练迭代次数设为 30，最大动作空间大小均设为 250。强化学习中，使用双重奖励机制驱动强化学习进行策略更新，累积折扣系数为 0.99，利用训练好的策略在知识图谱上搜索目标实体，并将最高奖励项目推荐至用户作为目标。由于演员－评论家网络是一个双层网络，设两层的大小分别为 512 和 256，损失函数由 L_r、$L_{\text{行为者}}$、$L_{\text{评论者}}$、L_{entropy} 组成，网络采用 Xavier[24] 进行初始化，优化函数为 Adam[24]。实验中，将前 10 个最佳结果作为准确率的比较。

4.2 基准方法和评价指标

将所提方法与 5 种基准推荐方法进行比较：

1）BPR[5] 是一种基于用户学习兴趣排序的贝叶斯推荐模型，利用用户－项目交互的嵌入形式实施推荐，而 KGR _ RL 提供了大量的辅助特征信息。

2）CF[7] 通过整合矩阵分解和异构数据学习用户兴趣，从而达到精准推荐的目标。CF 和 KGR _ RL 都使用了丰富的辅助信息，通过两种方法的比较说明深度强化学习结合知识图谱的有效性。

3）PGPR[8] 是一种基于知识图谱和元路径的可解释性推荐学习框架，它使用演员－评论家网络来寻找和推荐路径。PGPR 和 KGR _ RL 都是无监督的可解释推荐算法，其结果反映了可解释性的重要性。

4）JRL[9] 是一种联合表征学习模型，它将图像、文本和评级等多模态信息应用于神经网络。JRL 采用多模态信息表示，获得反映强化学习性能的精度结果。

5）DeepConn[10] 是一种基于深度强化学习的图卷积网络推荐方法，利用用户评论对产品信息进行编码提供有效的产品推荐。DeepConn 和 KGR _ RL 都使用了产品评论信息，通过两种方法的比较说明深度强化学习和辅助信息融合的有效性。

采用精度（Precision）、召回率（Recall）、命中率（HR，hit ratio）、归一化折损累积收益（NDCG，normalized discounted cumulative gain）个评估指标[25] 对所提方法的推荐性能进行评估，计算为：

$$Precision@K = \frac{1}{m}\sum_{u=1}^{m}\frac{hit_u}{K} \tag{18}$$

$$Recall@K = \frac{1}{m}\sum_{u=1}^{m}\frac{hit_u}{N_u} \tag{19}$$

$$HR@K = \frac{\sum_{u=1}^{m} hit_u}{\sum_{u=1}^{m} N_u} \tag{20}$$

$$NDCG@K = \frac{1}{IDCG@K}\sum_{i=1}^{K}\frac{2^{rel_i}-1}{\log_2(i+1)} \tag{21}$$

其中：m 为用户数，N_u 为用户 u 评价的物品数量；$IDCG@K$ 为用户最佳推荐结果 top-k 列表，rel 为推荐列表中排名第 i 位物品的分级相关性。

4.3　实验结果分析

4.3.1　性能比较

在“海洋测绘设备”和“工程测量设备”两个数据集上，对所提 KGR _ RL 与 5 种基准方法的推荐性能进行比较。如表 3 所示，KGR _ RL 模型在召回率、精准度、命中率和归一化折现累积收益等指标方面均优于其它 5 种基准方法。例如：KGR _ RL 的召回率指标比 PGPR 分别高 6.77%和 4.23%，精确率分别高 2.7%和 5.4%，归一化折现累积收益分别高 15.6%和 22.6%，主要是由于 KGR _ RL 采用了双重奖励机制和自监督损失约束作为增强信更新策略梯度，提高了整体推荐性能。

同时，通过 KGR _ RL 的两个变体进行消融实验，验证所提方法对推荐性能的影响：1）KGR _ RL-DR：采用双重奖励机制，不使用损失约束作为加强信号来监督策略梯度更新；2）KGR _ RL-L：使用损失约束作为加强信号来监督策略梯度更新，忽略双重奖励机制。从表 3 中可知，KGR _ RL-DR 和 KGR _ RL-L 在 4 个评价指标的推荐精度均优于其它 5 种基准方法，这说明将双重奖励机制和损失约束作为强化信号监督策略梯度更新的方法，可以有效提高模型质量和推荐精度。

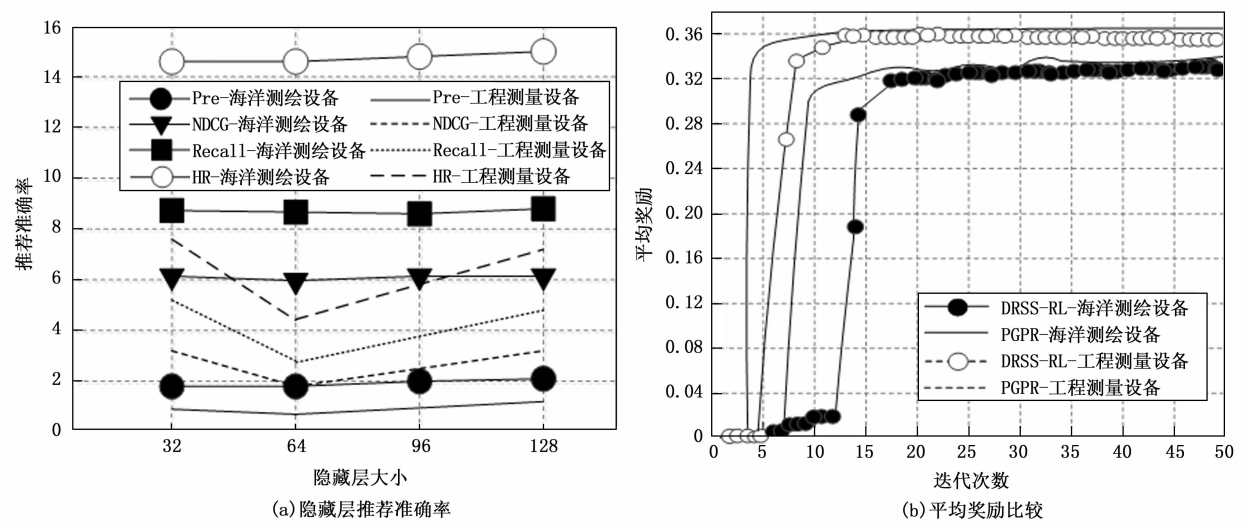
4.3.2　准确性和平均奖励分析

如图 3（a）所示，比较 KGR _ RL 在 3 个评价指标下的推荐准确性，当隐藏层大小为 64 时，KGR _ RL 表现最差，尤其是在“海洋测绘设备”数据集上，相比“工程测量设备”数据集的准确率更低。主要是由于模型在隐藏层为 64 时，在“工程测量设备”数据集上过度拟合。如图 3（b）所示，比较 KGR _ RL 与 PGPR 获得的平均奖励，虽然 PGPR 在后续训练迭代的平均奖励高于 KGR _ RL，但 KGR _ RL 的精准性优于 PGPR。平均奖励与推荐准确率的不一致反映了 PGPR 策略在选择行为和评估路径奖励时具有伪高的奖励，如果模型依赖这种伪高回报驱动策略进行行为选择，将获得高回报低准确率的状态。KGR _ RL 利用当前状态和后续状态评估计算策略选择行为反馈，利用增强损失约束纠正状态价值评估中存在误差的值函数，提高了模型策略精度，也不会从生成路径中获得总奖励的伪高值。

如图 4 所示，在两个勘测产品交易数据集上随着训

表 3　推荐性能比较

模型	海洋测绘设备				工程测量设备			
评价指标	*Recall*	*Precision*	*hit*	*NDCG*	*Recall*	*Precision*	*hit*	*NDCG*
BPR	1.038	0.185	1.171	0.621	4.189	1.134	8.256	2.771
CF	2.606	0.386	4.291	1.499	5.943	1.382	11.042	3.695
PGPR	4.874	0.738	6.937	2.853	8.442	1.749	14.552	5.528
JRL	2.898	0.429	4.629	1.732	6.952	1.612	12.891	4.913
DeepCoNN	2.328	0.242	3.289	1.135	5.941	1.118	9.769	3.371
KGR_RL-DR	5.061	0.758	7.292	2.979	8.590	1.757	14.756	5.631
KGR_RL-L	4.955	0.739	7.158	2.933	8.546	1.766	14.076	5.623
KGR_RL	5.082	0.765	7.377	3.009	8.792	1.803	15.028	5.754



(a)隐藏层推荐准确率　　(b)平均奖励比较

图 3　准确性和平均奖励分析

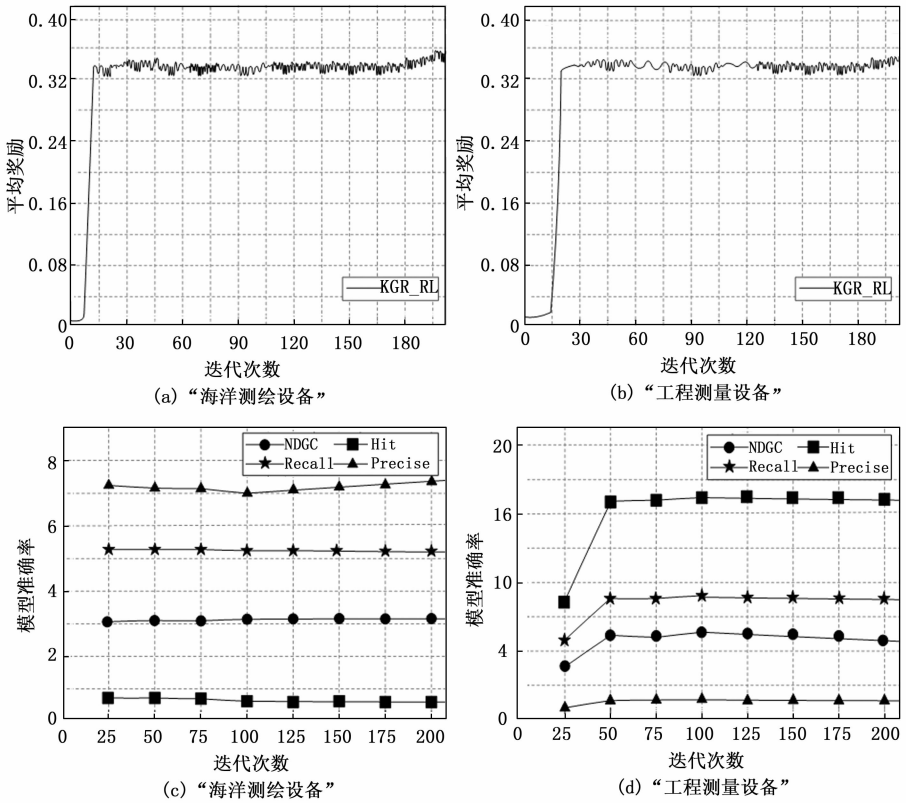


图 4 平均奖励比较

练次数增加平均奖励的数值变化情况，当迭代次数为 40 时，平均奖励和模型准确性逐渐稳定。主要是由于 KGR _ RL 在训练迭代早期存在欠拟合现象，平均奖励较低，模型的准确率较低，而后期逐步完成模型拟合后，平均奖励和准确率逐步提升并趋于稳定。

4.3.3 可解释性分析

可解释性在推荐时生成合理有效的推理路径，如表 4 所示，在数据集上比较了 KGR _ RL 与 KGR _ RL-DR、KGR _ RL-L 和 PGPR 等 3 种不同可解释推荐模型在召回率和精确度方面的性能。实验结果表明：KGR _ RL 的准确率和召回率均优于其它基准方法。例如：在“工程测量设备”数据集上，KGR _ RL 的召回率和精准度比 PGPR 分别提高了 2.5%、2.4%；在“海洋测绘设备”数据集上，KGR _ RL 的召回率和精准度比 PGPR 分别提高了 4.2%、1.5%，这是由于双重奖励可以引导策略寻找有效的推荐路径，利用增强损失函数更新路径推理策略，提高了训练数据集上的评价指标。

表 4 两个数据集上可解释性分析

数据集	海洋测绘设备		工程测量设备	
评价指标	Recall	Precision	Recall	Precision
PGPR	4.698	13.354	4.165	11.917
KGR_RL-DR	4.766	13.412	4.216	12.026
KGR_RL-L	4.746	13.398	4.20	11.963
KGR_RL	4.899	13.56	4.28	12.205
Lmp/%	4.2	1.6	2.5	2.4

通过双重奖励驱动机制展示了强化学习动作选择的策略优化，有效地调整了推理路径的数量和比例，并可获得比真实情景更多的奖励，这也与强化学习的目标一致。如表 5 所示，描述了通过一阶和二阶动作区分的不同推理路径的分布，推理路径分别由 PGPR 和 KGR _ RL 生成。与 PGPR 相比，KGR _ RL 生成的一阶行为

表 5 两个数据集上的有效性分析

数据	工程测量设备			海洋测绘设备		
方法	行为数	步骤 1 行为/比率	步骤 2 行为/比率	行为数	步骤 1 行为/比率	步骤 2 行为/比率
PGPR	1 329 736	Mention (97.89%)	Described_by(99.62%)	2 139 832	Mention (92.21%)	Described_by(98.6%)
			Mention (0.19%)			Mention (1.7%)
		Purchase (2.12%)	—		Purchase (7.88%)	—
KGR _RL	1 360 722	Mention (97.84%)	Described_by(99.99%)	2 342 269	Mention (90.86%)	Described_by (96.65%)
			Mention (0.02%)			Mention (0.36%)
		Purchase (2.15%)	—		Purchase (8.36%)	—

为“购买 Purchase”的推理路径比例增加，分别为 2 528和 25 185，而一阶行为“提及 Mention”的推理路径比例减少。此外，KGR _ RL 生成的推理路径中“描述 Described _ by”的二阶动作比例高于 PGPR 生成的推理路径。这是由于 KGR _ RL 模型不仅根据当前状态下行为空间中动作的相关性进行决策，还考虑了未来状态下行为选择的影响。

5　结束语

本文考虑产品环境感知状态选择对后续行为的影响，提出了一种基于强化学习的知识图谱产品推荐方法，主要贡献是结合短期奖励与长期增量评价提出双重奖励驱动策略，利用损失约束作为强化信号监督梯度更新推理策略，并以勘测交易产品为实验数据验证模型的有效性。但当数据规模较大时，演员－评论家网络可能会造成潜在信息丢失，下步将运用因果推理机制研究 KGR _ RL 的可解释性，并优化相关算法，以提高演员－评论家网络的稳定性。

参考文献：

[1] CHEN J Y, YU J, LU W J, et al. IR-Rec: an interpretive rules-guided recommendation over knowledge graph [J]. Information Sciences, 2021, 563: 326 - 341.

[2] TAO S H, QIU R H, XU B, et al. Micro-behaviour with reinforcement knowledge-aware reasoning for explainable recommendation [J]. Knowledge-Based Systems, 2022, 251: 109300.

[3] CHANG C, ZHOU J M, WENG Y, et al. KGTN: knowledge graph transformer network for explainable multi-category item recommendation [J]. Knowledge-Based Systems, 2023, 110: 48 - 54.

[4] WANG T X, ZHENG X L, HE S K, et al. Learning user-item paths for explainable recommendation [J]. IFAC-PapersOnLine, 2020, 53 (5): 436 - 440.

[5] SHAMS B, HARATIZADEH S. Graph-based collaborative ranking [J]. Expert Systems with Applications, 2017, 67: 59 - 70.

[6] LI Y K, WU Q, HOU L, et al. Entity knowledge transfer-oriented dual-target cross-domain recommendations [J]. Expert Systems with Applications, 2022, 195: 1165 - 1191.

[7] ALHARBE N, RAKROUKI M A. A collaborative filtering recommendation algorithm based on embedding representation [J]. Expert Systems with Applications, 2023, 215 (119): 73 - 83.

[8] BALLOCCU G, BORATTO L, FENU G, et al. Reinforcement recommendation reasoning through knowledge graphs for explanation path quality [J]. Knowledge-Based Systems, 2023, 260: 11009 - 11017.

[9] DU X Q, XUE Z G. JLCRB: A unified multi-view-based joint representation learning for CircRNA binding sites prediction [J]. Journal of Biomedical Informatics, 2022, 136: 1042 - 1073.

[10] WANG L P, ZHOU W, LIU L, et al. Deep adaptive collaborative graph neural network for social recommendation [J]. Expert Systems with Applications, 2023, 229 (A): 1204 - 1214.

[11] ZHU B, WU M, HONG Y P, et al MMIEA: Multi-modal interaction entity alignment model for knowledge graphs [J]. Information Fusion, 2023, 101: 9 - 54.

[12] ROBERTO G, JUAN-MIGUEL L, ROSA G. Rhizomer: Interactive semantic knowledge graphs exploration [J]. SoftwareX, 2022, 20: 1012 - 1047.

[13] XU X L, HONG Q F, WANG Y X, et al. A random walk sampling on knowledge graphs for semantic-oriented statistical tasks [J]. Data & Knowledge Engineering, 2022, 140: 1020 - 1044.

[14] PAUL A, SUAREZ-VARELA J, KRZYSZTOF R, et al. Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case [J]. Computer Communications, 2022, 196: 184 - 194.

[15] TOUTOU O, HAMZA G, SAMIR B. An agent-based model for resource provisioning and task scheduling in cloud computing using DRL [J]. Procedia Computer Science, 2021, 192: 3795 - 3804.

[16] TOKE T, FLEMMING S, CHRISTIPHER J. Interpreting outputs of agent-based models using abundance-occupancy relationships [J]. Ecological Indicators, 2012, 20: 221 - 227.

[17] LU F Y, ZHOU J, HUANG X L. Enhancing the convolution-based knowledge graph embeddings by increasing dimension-wise interactions [J]. Data & Knowledge Engineering, 2023, 146: 1021 - 1084.

[18] BAI L Y, CHAI D, ZHU L. RLAT: Multi-hop temporal knowledge graph reasoning based on reinforcement learning and attention mechanism [J]. Knowledge-Based Systems, 2023, 269: 1105 - 1119.

[19] ZHAO K, WANG X T, ZHANG Y R, et al. Leveraging demonstrations for reinforcement recommendation reasoning over knowledge graphs [C] //Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20). Association for Computing Machinery, New York, 2020: 239 - 248.

（下转第 282 页）