arXiv:2301.05944v1 [cs.IR] 14 Jan 2023

知识即力量，理解即影响：路径推理推荐中的效用及其他目标、解释质量与公平性

Giacomo Balloccu[1 [0000−0002−6857−7709]]、Ludovico Boratto[1 [0000−0002−6053−3015]]、Christian Cancedda[2 [0000−0002−8206−3181]]、Gianni Fenu[1 [0000−0003−4668−2476]]和 MirkoMarras[1 (✉) [0000−0003−1989−6057]]

[1] 卡利亚里大学，卡利亚里，意大利
giacomo.balloccu@acm.org，ludovico.boratto@acm.org
fenu@unica.it，mirko.marras@acm.org[2] 都灵理工大学，都灵，意大利 christian.cancedda@studenti.polito.it

摘要。 路径推理是一种重要的推荐方法，它基于知识图谱（KG）对高阶用户-产品关系进行建模。该方法可以提取推荐产品与用户已体验产品之间的推理路径，并将这些路径转化为面向用户的文本解释。然而，该领域的评估方案似乎存在异质性且较为有限，因此难以明确现有方法的影响。在本文中，我们复现了三种在顶级会议上提出的最先进的相关路径推理推荐方法。在基于两个公共数据集的统一评估方案下，并与其他知识感知方法进行比较，我们研究了这些方法在多大程度上满足推荐效用及超越目标、解释质量，以及消费者和提供者公平性。我们的研究呈现了该领域的进展，并指出了尚待解决的问题和未来方向。源代码：https://github.com/giacoballoccu/rep-path-reasoning-recsys。

关键词： 推荐系统、知识图谱、可复现性。

1 引言

推荐系统（RS）是实现个性化用户体验的一种常见策略 [29]。历史数据（例如浏览活动和评分）以及产品特征（例如标题和描述）是公认的 RS 训练数据来源。产品信息通常会通过知识图谱（KGs） [8, 27]加以补充。这些知识图谱包含实体 （例如用户、电影、演员）以及实体之间的关系 （例如某位演员出演了一部电影）。将知识图谱整合到推荐系统中，能够提升推荐效用 [34, 36], ，尤其是在稀疏数据和冷启动场景下 [18]。纳入知识图谱对于使推荐系统具备可解释性，并使推荐成为一种更透明的社会过程至关重要 [33, 44]。

基于知识图谱的知名推荐方法包括路径推理方法 [1, 23, 25, 26, 31, 37, 38, 41, 45]。为了指导推荐系统训练，这些方法依赖于知识图谱中刻画用户与产品之间高阶关系的路径，并从用户已体验过的产品与待推荐产品之间识别出相关路径。这些路径还可用于通过解释模板或文本生成来创建解释。在电影领域，路径“用户$_1$ 观看了电影$_1$ 导演为导演$_1$ 执导了$^{-1}$ 电影$_2$”可能会生成基于模板的解释：“向你推荐电影$_2$ ，是因为你看过电影$_1$ ，该片也由导演$_1$执导。”路径推理方法与正则化方法形成对比；后者根据产品特征对特定推荐的重要性为其赋予权重，但不提供任何解释 [16, 19, 34–36, 43]。

为推荐任务提出了大量知识图谱，并结合路径推理方法，以生成推荐结果和解释 [5]。然而，评估协议各不相同（例如，训练测试划分不同），且仅限于范围较窄的评估数据集和评估指标。以往研究通常表明，与（非）知识感知基线相比，新方法能够带来更高的推荐效用。但没有研究深入分析超越效用目标的表现（例如，覆盖率、意外发现性），也没有监测消费者（即终端用户）和提供者的公平性。因此，路径推理方法是否会在迄今尚未探索的目标之间强调某种权衡，仍不清楚。鉴于这一领域错综复杂，且研究视角过度聚焦于效用，我们需要一个通用的评估基准，以了解每种方法何时以及如何适用。

本文针对与路径推理方法相关、尚未探索的评估视角开展了一项可复现性研究（由不同团队在不同实验设置下进行）。首先，我们检索了顶级会议和期刊的论文集，找出了七篇相关论文。我们尝试依据已发布的源代码复现原始方法，但只有三种方法可以复现。其次，我们制定了一套通用评估方案，涵盖两个公共数据集（电影、音乐）、两个敏感属性（性别、年龄），以及涉及四个视角（推荐效用、超越效用目标、解释质量、公平性）的十六项评估指标。我们按照该方案评估了路径推理方法，并将其与其他知识感知方法进行了比较。结果表明，尽管这些方法的效用往往相近，但它们在实现其他推荐目标方面存在差异。本研究呼吁对这些方法开展更广泛的评估，并以更负责任的方式加以应用。

2 研究方法

本节介绍路径推理方法的收集过程、复现步骤以及通用评估方案。

2.1 论文收集

为收集现有的路径推理方法，我们系统地检索了顶级信息检索会议（CIKM、ECIR、ECML-PKDD、FAccT、KDD、RecSys、SIGIR、WSDM、WWW、UMAP）近期的论文集及顶级出版社（ACM、爱思唯尔、IEEE、Springer）出版的期刊。

表1. 本研究认为相关的路径推理方法。

方法	Year	状态[1]	实验设置				
			数据集[2]	划分比例[3]	划分方法[4]	推荐[5]	解释[5]
PGPR [41]	2019	RE	AZ	70-00-30	$Rand$	归一化折损累计增益，召回率，命中率，精确率	-
EKAR [31]	2019	$\overline{RE}$	ML, LFM, DB	60-20-20	$Rand$	归一化折损累计增益，命中率	-
CAFE [42]	2020	RE	AZ	70-00-30	$Rand$	归一化折损累计增益，召回率，命中率，精确率	-
UCPR [32]	2021	RE	ML, AZ	60-20-20	$Rand$	归一化折损累计增益，召回率，命中率，精确率	PPC
MLR [39]	2022	$\overline{RE}$	AZ	70-00-30	$Rand$	归一化折损累计增益，召回率，命中率，精确率	-
PLM-Rec [13]	2022	$\overline{RE}$	AZ	60-20-20	$Time$	归一化折损累计增益，召回率，命中率，精确率	-
TAPR [47]	2022	$\overline{RE}$	AZ	60-10-30	$Rand$	归一化折损累计增益，召回率，命中率，精确率	-

[1] **状态** RE：可复现且可扩展；RE：可复现但不可扩展；RE：不可复现且不可扩展。[2] **数据集** AZ：亚马逊 [22]；ML：MovieLens 1M [15]；LFM：LastFM [30]；DB：数据库2014 [9]。[3] **划分比例** 分别表示训练集、验证集和测试集所占的数据百分比。[4] **划分方法**。$Rand$：基于随机；$Time$：基于时间。[5] **评估指标** R：召回率；HR：命中率 P：精确率；PPC：路径模式集中度

采用的关键词由两类词组成：一类是技术术语，包括“路径推理推荐系统”和“可解释推荐系统”；另一类是非技术术语，包括“可解释人工智能”和“知识增强人工智能 ”。若论文满足以下条件，我们便将其标记为相关论文：（a）研究推荐；（b）提出基于知识图谱的方法；（c）该方法能够生成推理路径。我们排除了其他领域或任务的论文，例如非个性化排名或仅涉及实体预测任务（不涉及任何推荐）的论文。我们也排除了无法生成推理路径的知识感知方法，不过会选取其中的一些代表方法进行比较。本研究共选出七篇相关论文（表1）。

我们尝试复现每篇相关论文中的方法，并尽可能依赖原始源代码。为了获取源代码，我们首先尝试在原论文和网络上搜索源代码仓库。作为最后的手段，我们给原作者发送了电子邮件。若能获得可正常运行的源代码版本，且只需稍作修改即可使其适用于另一数据集并提取推荐结果（和推理路径），我们就认为该方法可以复现。七篇相关论文中，有三篇可以通过合理的投入完成复现。至于无法复现的论文，其中三篇没有提供任何源代码[3]。另一篇则包含无法获取的外部依赖项 [20]。

2.2 方法复现

对于每篇相关论文，我们都分析了所提出方法的原理以及实验设置的特点，如表1所总结。

PGPR [41]（原始源代码：https://github.com/orcax/PGPR）基于这样一种理念：训练一个强化学习（RL）智能体来寻找路径。在训练期间，智能体从用户出发，并学习以较高奖励抵达正确的产品。在推理期间，智能体直接遍历至正确的产品以进行推荐，而无需枚举用户与产品之间的所有路径。原始实验在四个 AZ数据集 [22]上进行，并基于产品元数据和评论构建了一个知识图谱。

[3] 请注意，受我们发送给原作者的电子邮件影响，这些论文的源代码可能很快会在线发布。我们将把对它们的复现留待未来工作。

EKAR [31] （未提供原始源代码）将该任务建模为用户-物品-实体图上的马尔可夫决策过程，并使用深度强化学习来求解。用户-物品-实体图被视为环境，智能体从中获取一系列已访问的节点和边。策略网络根据编码后的状态，输出动作空间上的概率分布。最后，如果智能体成功找到目标用户在训练集中消费过的产品，就会获得正奖励。其新颖之处在于使用 LSTM 作为策略网络，并设计了一个能使训练保持稳定、鼓励智能体探索的奖励函数。只有这项研究纳入了来自三个不同领域的数据集：电影（ML1M）、音乐（LFM）和书籍（数据库）。

CAFE [42] （原始源代码：https://github.com/orcax/CAFE）遵循由粗到细范式。在粗粒度阶段，给定知识图谱后，系统会创建用户画像，以捕捉以用户为中心的模式。为了在用户画像的引导下进行多跳路径推理，推理器被拆解为一系列神经推理模块。随后，系统根据用户画像组合这些模块，以高效执行路径推理。原始实验遵循 PGPR 的实验设置（相同的数据集、数据划分和评估指标）。

UCPR [32] （原始源代码：https://github.com/johnnyjana730/UCPR/）引入了一种多视图结构，不仅利用局部序列推理信息，还纳入了用户需求组合视图。用户需求组合在预处理阶段构建，并通过多步重新聚焦进行更新，使路径选择过程具备适应性且更加有效。原始实验设置涵盖电影（ML1M）和电子商务领域（AZ）。这项研究是唯一一项评估解释质量属性的研究，即选定路径中的知识图谱关系类型在多大程度上存在差异。

MLR [39] （通过电子邮件共享源代码，但缺少外部依赖项）是另一种强化学习框架，它利用本体视图和实例视图知识图谱来建模多层次用户兴趣。通过微软概念图（MCG） [20], ，该方法构建了多个概念层级（例如，Prada 是一个意大利奢侈时尚品牌）。随后，强化学习智能体通过在这些层级之间导航来进行推理。作者提供了源代码，但 MCG 已不再在线，所提供的知识图谱转储仅涉及最初使用的 AZ数据集。

PLM-Rec [13] （原始源代码不可用）在给定知识图谱的情况下，会提取满足不同跳数约束的训练路径序列。通过利用语言特征与语义的增强，该方法获得了一系列训练数据序列。随后，使用基于Transformer的解码器训练自回归路径语言模型。该方法能够缓解以往方法在知识图谱连通性方面的召回偏差。实验仅使用了 AZ数据集。

TAPR [47] （原始源代码不可用）提出了另一种基于强化学习（RL）的路径推理方法，其特点是在奖励函数中加入时间项。该时间项引导智能体进行具有时间意识的搜索，以捕捉用户兴趣的近期趋势。原始结果（在AZ数据集上）显示，与PGPR相比，该方法的效用有所提升；不过，该模型采用随机划分进行评估，这并不适合时间感知模型。

表2. 两个所研究数据集的交互与知识信息。

交互	ML1M	LFM1M
用户	6,040	4,817
产品	2,984	12,492
交互	932,295	1,091,275
密度	0.05	0.01
性别（年龄）组	2 (7)	2 (7)

知识	ML1M	LFM1M
实体（类型）	13,804 (12)	17,492 (5)
关系（类型）	193,089 (11)	219,084 (4)
稀疏性	0.0060	0.0035
总体平均度	28.07	25.05
产品平均度	64.86	17.53

2.3 评估方案

为确保不同方法之间的评估一致且统一，鉴于原始实验设置各不相同，我们结合了复现与再现 [7], ，但为方便起见，本文通篇仅使用“可复现性”一词。具体而言，我们依托原作者提供的源代码运行其方法，并使用自有数据和源代码来（a）按照各方法的要求预处理输入数据集；（b）根据其返回的推荐结果和推理路径计算评估指标。

数据收集。我们在两个数据集上进行了实验：MovieLens（ML1M） [15] 和 LastFM（LFM1B） [30]。鉴于我们关注公平性视角，我们选择了能够提供（或可以收集到）用户和提供方人口统计属性的数据集。因此，我们舍弃了其他数据集，例如亚马逊数据集 [22], ，因为无法合理地获取这些信息。所选数据集均为公开数据集，在领域、规模和稀疏性方面各不相同，从而在共同基准下，为复现的路径推理方法的泛化能力提供了新的见解，并可与其原始设置进行比较（见表1）。对于 ML1M，我们使用了 [9] 基于 DBpedia 生成的知识图谱；对于 LFM1B，我们则使用 [46] 从 Freebase 转储文件中提取的数据生成了知识图谱。

数据预处理。 对于 ML1M 数据集， [15]最初提供了消费者的性别和年龄敏感属性，但未提供提供方的相关属性。鉴于先前工作 [6],将导演视为电影提供方，我们采用了该研究中收集的导演敏感属性标签。在 LFM1B 中，只有一小部分终端用户附有性别和年龄标签。因此，我们剔除了所有无法获取敏感属性的用户。鉴于原始论文所用数据集均远小于 LFM1B，且我们的初步实验发现这些方法[4], 的可扩展性较低，因此我们从筛选后的 LFM1B 中均匀抽样，确保用户（产品）至少有 20（10）次交互。本文及相关结果中，我们将这一数据子集称为 LFM1M。该抽样使我们获得了与 ML1M 规模相当的数据集，并避免了我们不关注的冷启动场景。由于原始数据集未附带提供方的敏感属性（在音乐推荐系统中，艺术家通常被视为提供方），我们从 Freebase 中爬取了这些属性，并随本研究一并发布。

两个知识图谱均按照 [2] 中的方法进行了预处理，以统一三元组的格式。具体而言，我们仅考虑头实体为产品、尾实体为外部实体的三元组，从而为分析知识感知方法和基于路径的方法提供共同的数据集。

[4] 解决重大的可扩展性问题超出了我们可复现性研究的范围。

若考虑头实体和尾实体为外部实体或产品的三元组，则与复现研究相比，还需要构建额外的元路径（基于路径的方法所需），这超出了本研究的范围。此外，为控制稀疏性，我们移除了类型所对应三元组数量不足总数 3% 的关系。对于用户与产品之间的交互，我们剔除了知识图谱中不存在的产品（及其交互）。预处理后的数据集统计信息汇总于表2。

数据准备与划分。 对于每个数据集，我们首先按时间顺序对每位用户的交互进行排序。随后，我们采用基于时间的留出策略进行训练-验证-测试划分：最早的 60% 交互作为训练集，接下来的 20% 用于验证，最近的 20% 作为测试集。上述预处理后的数据集用于训练、优化和测试每个基准模型。这样，我们便能在真实场景中开展评估，在这种场景下，可能决定交互模式的趋势并非固定不变，而是会随时间演变。

知识感知模型比较。 路径推理方法属于知识感知推荐这一类中的一个子类。为了更好地说明本研究的背景，我们因此决定与两个基于知识嵌入的知识感知模型（CKE [43] 和 CFKG [1], ），以及一个基于传播的知识感知模型 KGAT [36] 进行比较（在比较有意义时）。这三个模型无法向用户提供推理路径，我们按照相同的协议[5]对其进行了复现和评估。为简洁起见，我们不详细说明其复现过程，读者可参阅我们的代码仓库。

超参数微调。 给定一个数据集和一个模型，我们通过网格搜索选择最佳超参数设置。该网格搜索涵盖了原论文中发现较为敏感的超参数及其取值。在某些情况下，根据初步实验的结果，我们扩展了取值网格，以更好地适应当前数据集的特征。关于超参数及其在网格搜索中的取值的完整细节，已在我们的代码库中报告。我们在验证集上评估通过不同超参数设置获得的模型，并选择归一化折损累计增益最高的模型。

评估指标计算。 给定一个模型和一个数据集，我们基于相应的测试集，在大小为 $k = 10$ （例如， [7]）的推荐列表上，监测推荐效用、超越效用目标、解释质量，以及消费者与提供者公平性。我们在表 3 中描述了每项指标，并在代码库中提供了实现细节。

对于消费者的推荐效用，我们使用二元相关性分数和以 2 为底的对数衰减，监测了归一化折损累计增益（NDCG） [40], 和平均倒数排名（MRR） [11]。与召回率和准确率不同，NDCG 会考虑相关产品在推荐列表中的位置。平均倒数排名则只考虑首个相关产品的位置，因此提供了不同于 NDCG 的视角。

[5] 为简洁起见，我们未纳入非知识感知方法（例如贝叶斯个性化排序），尽管在 [4] 等研究中，这些方法在某些评估指标（例如归一化折损累计增益）上与路径推理方法进行了比较。尽管如此，这仍是未来研究的一个重要方向。

表 3. 可复现性研究中涵盖的评估指标。

视角	指标	缩写	范围	描述
消费者效用	归一化折损累计增益	NDCG	[0, 1]	推荐产品的对用户有用的程度（1 表示最有用）。
	平均倒数排名	MRR	[0, 1]	首个推荐产品对用户的效用程度（1 表示更有用）。
消费者超越效用	覆盖率	COV	(0, 1)	产品总数的百分比至少被推荐过一次（1 表示覆盖率高）。
	多样性	DIV	(0, 1]	涵盖的产品类别百分比在推荐列表中（1 表示多样性高）。
	新颖性	NOV	(0, 1)	推荐产品受欢迎程度的倒数给用户（1 表示流行度低，因此新颖性高）。
	意外发现性	SER	[0, 1]	推荐产品中不.....基线也推荐的产品所占的百分比（数值为 1 表示更加出乎意料）。
消费者解释质量	忠实度	FID	[0, 1]	可解释的推荐产品所占的比例（1 表示所有产品都可解释）。
	关联交互时效性	LIR	[0, 1]	路径中既往交互的新近性以及相应推荐产品的交互（1 表示较新）。
	关联交互多样性 LID		(0, 1]	路径中不同过去交互的数量这些交互伴随的推荐产品（1 表示不同）。
	共享实体流行度	SEP	[0, 1]	路径中共享实体的流行度伴随推荐产品的数量（1 表示热门）。
	共享实体多样性	SED	(0, 1]	推荐产品所关联路径中的不同共享实体数量伴随推荐产品的数量（1 表示不同）。
	路径类型多样性	PTD	(0, 1]	不同路径类型在路径中的占比所伴随的推荐产品（1 表示不同）。
	路径类型集中度	PTC	[0, 1]	不同路径类型表示是否均衡（1 表示均衡）。
提供方效用	曝光度	EXP	[0, 1]	给定提供方的项目在推荐列表中的曝光度（1 表示曝光度高）。

在我们的研究中，我们还关注了四个著名的超越效用目标 [21]。我们监测生成的推荐结果对可用产品目录的覆盖程度（覆盖率）。较高的覆盖率可能会提升用户的满意度和销售额。另一项目标是多样性，它被认为与人类理解 [12]和内容接受度 [10]相关。我们将其计算为推荐列表中不同产品类别所占的百分比。此外，意外发现性衡量推荐惊喜度 [17]。鉴于我们采用离线设置，我们将推荐结果与基线模型（即推荐最流行产品的系统） [24]进行比较。基准模型与基线模型的推荐结果差异越大，意外发现性就越高。最后，我们将新颖性估算为产品流行度的倒数（根据收到的评分计算），并假设流行度较低的产品更可能带来惊喜 [48]。

关于解释质量，我们考察了推荐列表中可解释产品的比例（忠实度） [28]。此外，我们监测了推理路径在新近性、流行度和多样性方面的属性 [3]。较新的关联交互有助于用户更好地理解解释。关联交互新近度衡量解释路径中所呈现过去交互的新近程度，而关联交互多样性则监测其中包含多少个不同的过去交互。第二个视角与共享实体有关，其假设是：流行度更高的共享实体更有可能为用户所熟悉。共享实体流行度衡量解释路径中共享实体的流行度（即其在知识图谱中的节点度）。相反，共享实体多样性则监测不同共享实体的数量。最后，路径类型多样性关注其中包含多少种不同的路径类型。路径类型集中度则监测各路径类型的分布是否均衡。

随着公平性日益受到重视，我们也根据人口统计均等这一概念评估了公平性 [7,14]。对于消费者公平性，给定一个指标，我们计算了各人口统计群体中该指标的平均值，并监测群体之间的绝对两两差异[6]。关于提供者公平性，我们计算了特定人口统计群体中各提供方产品获得的平均曝光度 [6]。最后，我们同样计算了提供方群体之间的平均绝对两两差异。

3 实验结果

我们的研究旨在通过回答以下研究问题，探究路径推理方法的多个评估视角：

RQ1 路径推理方法是否会为了提升解释能力而牺牲推荐效用和/或超越效用目标？**RQ2** 路径推理方法在多大程度上能够为所有推荐产品生成解释，这取决于推荐列表的大小吗？**RQ3** 基于路径类型和特征，不同路径推理方法所选路径的质量有何差异？

3.1 为解释能力而权衡推荐目标（RQ1）

首先，我们分析了所考察的路径推理方法（PGPR、CAFE、UCPR）与相关的、利用知识但不具备可解释性的方法（KGAT、CKE、CFKG）之间，在推荐效用和超越效用目标方面是否存在显著差异。为评估统计显著性，我们针对每个数据集，将两类方法视为两个独立组，并对每项指标进行t检验。这样也使我们能够考察知识图谱稀疏性对表现的影响。尽管总样本量（六种方法）相当小，但这种设置仍有助于发现两类方法的一些初步特征。不过，仍需针对更广泛的方法集开展进一步研究，以更充分地评估结果的泛化能力。每项测试得到的P值见表5。

图1展示了我们在效用（归一化折损累计增益、最大边际相关性）和超越效用目标（意外发现性、多样性、新颖性和覆盖率）方面的评估结果。我们还报告了提供者公平性的估算结果，而消费者公平性将在后文讨论。尽管图1便于比较，但具体数值请参见表4。就推荐效用而言，在 ML1M 数据集上，路径推理方法取得的得分（归一化折损累计增益为 0.26 至 0.28；平均倒数排名为 0.18 至 0.21）与知识感知的不可解释方法（归一化折损累计增益为 0.26 至 0.29；平均倒数排名为 0.21 至 0.23）相当。但在 LFM1M 数据集上，情况并非如此：路径推理方法产生的推荐效用（归一化折损累计增益为 0.15 至 0.34；平均倒数排名为 0.09 至 0.27）低于知识感知基线（归一化折损累计增益为 0.13 至 0.40；平均倒数排名为 0.10 至 0.34）。不过，在两种情况下、两个数据集上，不同方法类别在推荐效用方面均未发现统计学差异（所有 P 值均大于 0。05）。

[6] 根据数据集标签，我们的数据包含与性别（男性、女性）和年龄（18岁以下、18至24岁、25至34岁、35至44岁、45至49岁、50至55岁、 56+）相关的敏感属性。

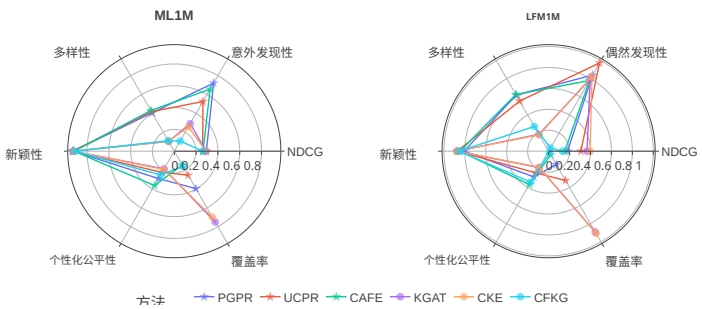


图1. 推荐效用及超越效用目标的比较 [RQ1]。

表4. 推荐效用及超越效用目标的指标得分 [RQ1]。

方法	ML1M							LFM1M						
	归一化折损累计增益 ↑	最大边际相关性 ↑	偶然性 ↑	多样性 ↑	新颖性 ↑	PF[1] ↓₀	覆盖率 ↑	归一化折损累计增益 ↑	最大边际相关性 ↑	偶然性 ↑	多样性 ↑	新颖性 ↑	PF[1] ↓₀	覆盖率 ↑
CKE	**0.29**	**0.23**	0.26	0.10	**0.93**	0.19	0.70	**0.40**	**0.34**	0.82	0.18	0.88	0.18	**0.91**
CFKG	0.26	0.21	0.11	0.11	0.92	0.25	0.16	0.13	0.10	0.04	0.27	0.86	0.34	0.02
KGAT	**0.29**	**0.23**	0.29	0.10	**0.93**	0.19	**0.75**	0.37	0.31	0.79	0.19	**0.88**	0.18	0.89
PGPR	0.28	0.21	0.78	0.42	0.93	0.27	0.42	0.31	0.25	0.81	0.54	0.82	0.32	0.20
UCPR	0.26	0.20	0.53	0.42	**0.93**	0.22	0.25	0.34	0.27	**0.94**	0.57	0.87	0.22	0.41
CAFE	0.26	0.18	0.63	**0.44**	**0.93**	**0.36**	0.21	0.15	0.09	0.75	**0.58**	0.84	**0.36**	0.11

对于每个数据集：最佳结果以**粗体**显示，次佳结果加下划线。 [1] **评估指标** PF：提供者公平性。

表5. 通过t检验P值评估基于路径的方法与知识感知方法在各项评估指标上的统计显著性差异 [RQ1]。

数据集[1]	归一化折损累计增益	最大边际相关性	偶然性	多样性	新颖性	提供商公平	覆盖率
ML1M	0.33	0.08	**0.01**	**0**	0.42	0.21	0.33
LFM1M	0.77	0.65	0.38	**0**	0.16	0.38	0.34

低于 0.05 的 P值以**粗体**显示。

就超越效用目标而言，路径推理方法取得的意外发现性显著高于知识感知的不可解释基线（ML1M 上为 0.53 至 0.78，LFM1M 上为 0.75 至 0.94；后者在 ML1M 上为 0.11 至 0.29，在 LFM1M 上为 0.04 至 0.82）。有趣的是，结果证实路径推理方法在 LFM1M 上的表现往往不如在 ML1M 上。此外，统计检验显示，在 ML1M 上，两类方法在意外发现性方面表现不同（P 值为 0.011）；平均而言，路径推理方法的意外发现性高于知识感知方法。多样性方面也呈现类似趋势。在 ML1M 和 LFM1M 上，路径推理方法的平均多样性均较高（路径推理方法与知识感知方法分别为：LFM1M 上 0.56 和 0.21；ML1M 上 0.43 和 0.10）（P 值 ≈ 0）。相反，在覆盖率方面，表现最佳的路径推理方法比表现最佳的知识感知方法在 ML1M 上低 44%，在 LFM1M 上低 54.9%。这可能是由于知识图谱中可供路径推理方法使用的路径数量较少。除 CFKG 外，有证据表明，如果目标是优化覆盖率，应优先选择知识感知方法。最后，两类方法的新颖性得分相近。

从提供者公平性的视角来看，与另一类方法相比，路径推理方法使提供商群体的曝光度更加公平（ML1M 上为 0.22 至 0.36；LFM1M 上为 0.22 至 0.36），而另一类方法为（ML1M 上为 0.19 至 0.25；LFM1M 上为 0.18 至 0.34）。令人意外的是——

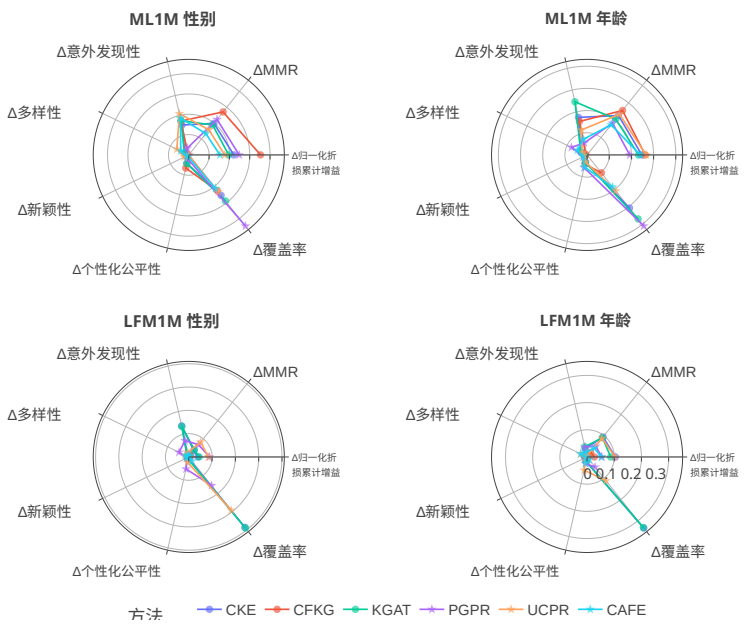


图2。 不同群体（性别和年龄）在推荐效用、超越效用目标和提供者公平性方面的差别影响。数值越低，公平性越高 [RQ1]。

——尽管推荐效用较低，CFKG 的提供者公平性得分仍位居第二。另一方面，根据所考察的评估指标得出的消费者公平性估计[7] 汇总于图 2 中。由于不同数据集和人口群体中的模式相似，我们仅介绍 ML1M 数据集和性别群体的结果。对于后者，所有模型在效用方面（NDCG 和 MRR）都表现出一定程度但较低的不公平性。路径推理方法（PGPR、UCPR、CAFE）在覆盖率方面的不公平性更高，分别为 0.084、0.045 和 0.032；相比之下，基线方法中 KGAT 的覆盖率不公平性最高，为 0.06。对于其他指标，我们发现差异虽小，但具有可比性。

RQ1的发现。 路径推理方法以推荐效用和覆盖率为代价换取解释能力，在*LFM1M*中尤为明显。相反，与知识感知的不可解释基线相比，这些方法在其他超越效用目标和提供者公平性方面取得了更高的估计值。

3.2 为所有推荐产品生成解释（RQ2）

在第二项分析中，我们希望了解路径推理方法在不同规模的推荐列表中，能够为所有推荐产品生成解释的程度。这一属性，即忠实度，对于

[7] 在平均得分方面表现最佳和最差的人口群体之间的差异，在t检验（如适用）或H检验（否则）下均具有统计显著性。

表 6. 不同截断点下的解释忠实度分析 $k = \{10, 20, 50, 100\}$ [RQ2]。

方法	ML1M				LFM1M			
	10	20	50	100	10	20	50	100
PGPR	1.00	0.99	0.99	0.78	0.98	0.74	0.31	0.15
CAFE	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
UCPR	0.61	0.34	0.14	0.07	0.99	0.98	0.68	0.35

能够生成推理路径（并提供解释）的方法至关重要。表 6 展示了两个数据集在推荐列表规模分别为 10、20、50 和 100 时的忠实度分数。

对于 PGPR，在推荐列表长度达到 50 之前，几乎每个产品都附有路径。当列表长度为 100 时，忠实度显著下降至 78%。在 LFM1M 数据集上，这种忠实度下降更为明显。在后一个数据集中，列表长度仅为 20 时，就只有 74% 的推荐产品获得了解释。对于 UCPR，我们在所考察的两个数据集上观察到与 PGPR 类似但相反的模式。在 ML1M 上，即使列表长度为 10，UCPR 也难以生成解释（只有 61% 的产品获得了解释）。相反，在 LASTFM 上，同一模型的忠实度高于 PGPR，列表长度为 10 时，获得解释的产品比例为 99%。令人意外的是，CAFE 能够为列表长度不超过 100 的每个推荐产品提供一条推理路径。不过需要注意的是，要实现这一点，必须在训练期间预先指定列表长度。CAFE 会根据列表长度自动调整周围待搜索邻域的大小。因此，当列表长度预先已知、固定或不超过某个上限时，这种方法将是最佳选择。尽管忠实度可以得到控制，但这样做会导致 CAFE 的归一化折损累计增益较低（见第 3.1 节）。

RQ2 的发现。 路径推理方法在忠实度方面呈现出截然不同的模式。*CAFE* 的忠实度较高，且在不同数据集和推荐列表规模下都保持稳定。另一方面，*PGPR* 在 *ML1M* 上的忠实度高于 *LFM1M*，但会迅速下降；*UCPR* 则正好相反。

3.3 解释质量的差异（RQ3）

最后，我们分析了选定路径（进而也包括由此生成的解释）的质量如何随路径特征而变化。为此，表 7 汇总了各推理路径方法和数据集的七个解释路径质量视角（LIR、LID、SEP、SED、PTD、PTC、PPC）。

就新近性维度而言，我们未观察到三种方法在关联交互新近度方面存在显著差异；在 ML1M 上，PGPR（0.44）取得最高值，CAFE（0.34）取得最低值（LFM1M 上的情况类似）。而在关联交互多样性方面，有趣的是，PGPR（ML1M 上为 0.84；LFM1M 上为 0.77）和 UCPR（ML1M 上为 0.82；LFM1M 上为 0.84）的多样性均高于 CAFE。

转向流行度视角，尤其是共享实体流行度方面，CAFE 在两个数据集上均取得了最高的 SEP（ML1M 为 0.75；LFM1M 为 0.77），这意味着它倾向于生成包含流行度更高的共享实体的路径。与 CAFE 相比，PGPR 和 UCPR 的数值则

表 7. 解释质量分析 [RQ3]。

模型	ML1M							LFM1M						
	LIR ↑	LID ↑	SEP ↑	SED ↑	PTD ↑	PTC ↑	PPC ↑	LIR ↑	LID ↑	SEP ↑	SED ↑	PTD ↑	PTC ↑	PPC ↑
PGPR	**0.44**	**0.84**	0.43	0.99	0.12	0.03	0.12	**0.49**	0.77	0.61	0.94	0.30	0.05	0.24
CAFE	0.34	0.16	**0.75**	**1.00**	**0.33**	**0.73**	**0.37**	**0.49**	0.25	**0.77**	**0.98**	0.25	**0.62**	**0.50**
UCPR	0.40	0.82	0.35	0.92	0.24	0.01	0.24	**0.49**	**0.84**	0.67	0.78	**0.42**	0.24	0.34

Fo对于每个数据集：最佳结果用**粗体**标出，次佳结果加下划线标出。

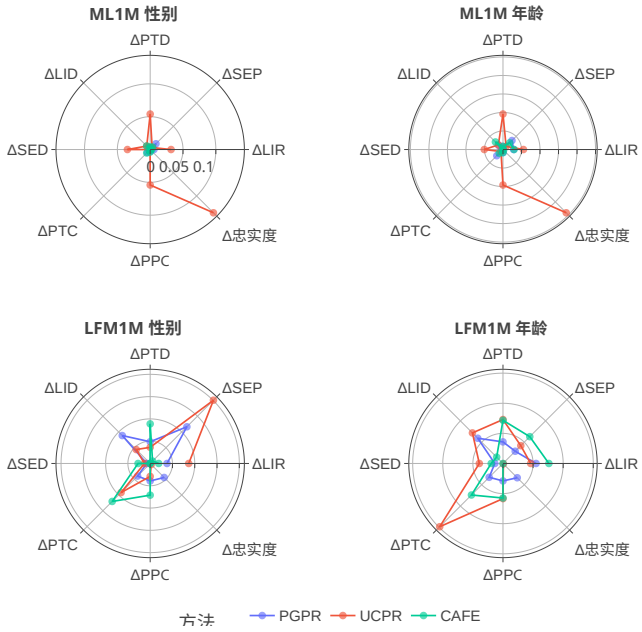


图 3. 不同群体（性别和年龄）在新近性、流行度和多样性方面的解释质量差别影响。数值越低，公平性越高 [RQ3]。

明显低得多。所有方法的 SED 估计值都很高（ML1M 为 0.92 至 1；LFM1M 为 0.78 至 0.98）。因此，这些方法能够在其推理路径中纳入种类丰富的共享实体。

关于路径类型的模式尤其有趣。具体而言，CAFE 和 UCPR 的路径类型多样性和路径类型集中度都高于 PGPR。这表明，前两者的解释平均而言在路径类型方面更加丰富（例如，主演、导演），而 PGPR 的解释平均而言则局限于一小组不同的路径类型。

图 3 展示了不同人口统计群体之间给定指标平均分的两两差异（消费者公平性）。同样，在 RQ1 的相同条件下，所有差异在统计上均有效。为简洁起见，我们仅讨论性别群体的结果。在两个数据集上，PGPR 和 CAFE 在所有评估指标上的不公平性估计都很低：UCPR 中所有分数均低于 0.01，PGPR 中均低于 0.02。相比之下，UCPR 在 PTD、SED 和 PPC 上表现出不公平性。值得注意的是，在两个数据集上，最明显的差别影响出现在 FID（0.12）、PTD 和 PPC（0.05），以及 SED（0.03）。

我们推测，这些不公平性估计可能是由不同人口群体之间的数据不平衡造成的。

RQ3 的发现。 路径推理方法在新近性、流行度和多样性方面，往往会产生差异显著的路径。尽管存在这些差异，但未发现对解释质量有显著的差别影响。

4 讨论与结论

本节将各项实验得出的主要发现联系起来，并阐述本可复现性研究的启示与局限。

在第一项分析中，我们比较了能够生成解释的方法（路径推理）与知识感知的不可解释基线在效用、超越效用目标以及提供者公平性方面的表现。结果显示，这两类方法在效用方面略有差异，尽管解释可能具有说服作用，而这种作用无法通过离线方式捕捉。未来研究应探讨解释如何影响用户决策，进而影响效用。对于超越准确率的目标，我们观察到，由于其内部机制，路径推理方法往往倾向于偶然性和多样性。同时，为生成解释而采用的方法设计使这些方法对知识图谱结构更加敏感，因而导致基准方法的产品覆盖率较低。结果还表明，所有方法（包括基线）在几乎所有视角下都表现出一定程度的不公平，尤其是在效用方面。我们的研究呼吁开发能够在知识感知设定下考虑多个视角的去偏方法。

在第二项分析中，我们分析了路径推理方法能否针对不同规模的推荐列表生成解释。结果表明，一些模型（如 UCPR）对数据和知识图谱的组成更为敏感，这会影响它们在推荐列表较短时生成推理路径的能力。通过进行特定的模型设计选择，可以避免这一局限。例如，CAFE 为每位用户采用预定义的搜索空间，为每个推荐产品提供推理路径（尽管这一设计选择可能会影响推荐效用）。

在最后一项分析中，我们不再仅关注生成推理路径的能力，而是进一步考察其质量。多项研究强调了解释的益处 [33]。近期研究还表明，路径属性（如新近性、流行度和多样性）会影响用户对解释的感知 [3]。结果显示，这些目标并非总能同时实现。例如，PGPR 无法在 ML1M 中生成多样化的解释，而 CAFE 生成的解释则仅基于少量过往用户交互。未来研究应通过处理前和处理后的方法解决这一问题，并考察其他解释视角（如说服力、信任和效率）。

总体而言，我们的分析表明，复现该领域的研究仍是一项艰巨的任务。在未来的工作中，我们计划详细探究知识图谱特征对所考察各个视角的影响，并设计能够适应知识图谱结构、且能有效兼顾多个目标的新型路径推理方法。

参考文献

1. Ai, Q., Azizi, V., Chen, X., Zhang, Y.: 面向可解释推荐的异构知识库嵌入学习。Algorithms **11**(9), 137 (2018) 2. Balloccu, G., Boratto, L., Fenu, G., Marras, M.: 利用知识图谱实践可解释推荐系统。载于：第16届ACM推荐系统会议论文集。第710–713页。RecSys '22，计算机协会，美国纽约州纽约市（2022）3. Balloccu, G., Boratto, L., Fenu, G., Marras, M.: 利用知识图谱对推荐系统进行后处理，以提升解释的新近性、流行度和多样性。载于：第45届国际ACM SIGIR信息检索研究与发展会议论文集。第646–656页。SIGIR '22，计算机协会，美国纽约州纽约市（2022）4. Balloccu, G., Boratto, L., Fenu, G., Marras, M.: 基于知识图谱的强化推荐推理，以提升解释路径质量。Knowledge-Based Systems **260**, 110098 (2023) 5. Barredo Arrieta, A., Díaz Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: 可解释人工智能（XAI）：迈向负责任人工智能的概念、分类体系、机遇与挑战。Inf. Fusion **58**, 82–115 (2020) 6. Boratto, L., Fenu, G., Marras, M.: 上采样与正则化在推荐系统提供者公平性中的相互作用。User Model. User Adapt. Interact.**31**(3), 421–455 (2021) 7. Boratto, L., Fenu, G., Marras, M., Medda, G.: 推荐系统中的消费者公平性：定义与缓解措施的语境化。载于：Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., Setty, V.（编）。信息检索进展。第552–566页。Springer, Cham（2022）8. Cao, Y., Hou, L., Li, J., Liu, Z.: 神经集体实体链接。载于：第27届国际计算语言学会议论文集。第675–686页。计算语言学协会，美国新墨西哥州圣菲（2018年8月）9. Cao, Y., Wang, X., He, X., Hu, Z., Tat-seng, C.: 统一知识图谱学习与推荐：加深对用户偏好的理解。载于：WWW（2019）10. Carbonell, J., Goldstein, J.: 使用最大边际相关性（MMR）进行基于多样性的重排序，以重新排列文档并生成摘要。载于：第21届年度国际ACM SIGIR信息检索研究与发展会议论文集。第335–336页。SIGIR '98，计算机协会，美国纽约州纽约市（1998）11. Craswell, N.: 平均倒数排名，第1703–1703页。Springer US（2009）12. Gedikli, F., Jannach, D., Ge, M.: 我该如何解释？推荐系统中不同解释类型的比较。International Journal of Human-Computer Studies **72**(4), 367–382 (2014) 13. Geng, S., Fu, Z., Tan, J., Ge, Y., de Melo, G., Zhang, Y.: 知识图谱上的路径语言建模，用于可解释推荐。载于：2022年ACM Web会议论文集。第946–955页。WWW '22，计算机协会，美国纽约州纽约市（2022）14. Gómez, E., Zhang, C.S., Boratto, L., Salamó, M., Marras, M.: 赢者通吃：教育推荐系统中的地理失衡与提供者（不）公平性。载于：SIGIR '21：第44届国际ACM SIGIR信息检索研究与发展会议，线上活动，加拿大，2021年7月11日至15日。第1808–1812页。ACM（2021）

15. Harper, F.M., Konstan, J.A.：MovieLens 数据集：历史与背景。ACM Trans. Interact. Intell. Syst. **5**(4)（2015年12月）16. He, G., Li, J., Zhao, W.X., Liu, P., Wen, J.R.：通过对抗学习，从用户-物品交互数据中挖掘隐式实体偏好以完成知识图谱。载于：2020年万维网大会论文集。第740–751页（2020）17. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.：评估协同过滤推荐系统。ACM Trans. Inf. Syst. **22**(1), 5–53（2004年1月）18. Huang, C., Gan, Z., Ye, F., Wang, P., Zhang, M.：KNCR：面向推荐系统的知识感知神经协同排序。载于：IEEE国际可靠、自主与安全计算会议、国际普适智能与计算会议、国际云计算与大数据计算会议、国际网络科学与技术大会，DASC/PiCom/CBDCom/CyberSciTech 2020，加拿大卡尔加里，2020年8月17日至22日。第339–344页。IEEE（2020）19. Huang, J., Zhao, W.X., Dou, H., Wen, J.R., Chang, E.Y.：利用知识增强记忆网络改进序列推荐。载于：第41届国际ACM SIGIR信息检索研究与开发会议。第505–514页。SIGIR '18，计算机协会，纽约，美国（2018）20. Ji, L., Wang, Y., Shi, B., Zhang, D., Wang, Z., Yan, J.：微软概念图：挖掘语义概念以理解短文本。《数据智能》**1**(3), 238–270（2019年6月）21. Kaminskas, M., Bridge, D.：多样性、意外发现性、新颖性和覆盖率：对推荐系统中超越准确性目标的综述与实证分析。ACM Trans. Interact. Intell. Syst. **7**(1)（2016年12月）22. Linden, G., Smith, B., York, J.：Amazon.com 推荐：物品到物品的协同过滤。IEEE Internet Comput. **7**(1), 76–80（2003）23. Ma, W., Zhang, M., Cao, Y., Jin, W., Wang, C., Liu, Y., Ma, S., Ren, X.：利用知识图谱联合学习可解释的推荐规则。载于：万维网大会论文集。第1210–1221页（2019）24. Murakami, T., Mori, K., Orihara, R.：评估推荐列表意外发现性的指标。载于：Satoh, K., Inokuchi, A., Nagao, K., Kawamura, T.（编）《人工智能新前沿》。第40–46页。施普林格柏林海德堡，德国柏林、海德堡（2008）25. Musto, C., de Gemmis, M., Lops, P., Semeraro, G.：为推荐系统生成基于评论的事后自然语言说明。User Modeling and User-Adapted I nteraction **31**(3), 629–673（2021）26. Ni, J., Li, J., McAuley, J.J.：利用远程标注评论和细粒度方面为推荐结果提供依据。载于：Inui, K., Jiang, J., Ng, V., Wan, X.（编）《2019年自然语言处理经验方法会议和第九届国际自然语言处理联合会议论文集》，EMNLP-IJCNLP 2019，中国香港，2019年11月3日至7日。第188–197页。计算语言学协会（2019）27. Oramas, S., Ostuni, V.C., Di Noia, T., Serra, X., Di Sciascio, E.：基于知识图谱的声音与音乐推荐。ACM Trans. Intell. Syst. Technol. **8**(2), 21:1–21:21（2017）28. Peake, G., Wang, J.：解释挖掘：推荐系统潜因子模型的事后可解释性。载于：第24届ACM SIGKDD知识发现与数据挖掘国际会议论文集。第2060–2069页。KDD '18，计算机协会，纽约，美国（2018）29. Ricci, F., Rokach, L., Shapira, B.：《推荐系统手册》，第1-35卷，第1–35页。Springer（2010年10月）

30. Schedl, M.：用于音乐检索与推荐的lfm-1b 数据集。载于：2016年ACM国际多媒体检索会议论文集。第103–110页。ICMR '16，ACM，纽约，美国（2016）31. Song, W., Duan, Z., Yang, Z., Zhu, H., Zhang, M., Tang, J.：EKAR：一种用于知识感知推荐的可解释方法。CoRR **abs/1906.09506**（2022）32. Tai, C.Y., Huang, L.Y., Huang, C.K., Ku, L.W.：面向可解释推荐的以用户为中心的路径推理，第879–889页。计算机协会，纽约，美国（2021）33. Tintarev, N., Masthoff, J.：推荐系统中解释的综述。载于：第23届国际数据工程会议研讨会论文集，ICDE 2007。第801–810页。IEEE计算机学会（2007）34. Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., Guo, M.：RippleNet：在知识图谱上传播用户偏好的推荐系统。载于：第27届ACM国际信息与知识管理会议论文集。第417–426页。CIKM '18，计算机协会，纽约，美国（2018）35. Wang, H., Zhang, F., Xie, X., Guo, M.：DKN：用于新闻推荐的深度知识感知网络（2018）36. Wang, X., He, X., Cao, Y., Liu, M., Chua, T.S.：KGAT：面向推荐的知识图谱注意力网络。载于：第25届ACM SIGKDD知识发现与数据挖掘国际会议论文集。第950–958页。KDD '19，计算机协会，纽约，美国（2019）37. Wang, X., Wang, D., Xu, C., He, X., Cao, Y., Chua, T.S.：基于知识图谱的可解释推荐推理。载于：第三十三届AAAI人工智能会议、第三十一届人工智能创新应用会议和第九届AAAI人工智能教育进展研讨会论文集。AAAI出版社（2019）38. Wang, X., Wang, D., Xu, C., He, X., Cao, Y., Chua, T.S.：基于知识图谱的可解释推荐推理。载于：第三十三届AAAI人工智能会议、第三十一届人工智能创新应用会议和第九届AAAI人工智能教育进展研讨会论文集。AAAI出版社（2019）39. Wang, X., Liu, K., Wang, D., Wu, L., Fu, Y., Xie, X.：基于强化学习的知识图谱多层级推荐推理。载于：ACM Web会议2022论文集。第2098–2108页。WWW '22，计算机协会，纽约，美国（2022）40. Wang, Y., Wang, L., Li, Y., He, D., Liu, T.：NDCG类排序指标的理论分析。载于：COLT 2013——第26届年度学习理论会议，2013年6月12日至14日，美国新泽西州普林斯顿。JMLR研讨会与会议论文集，第30卷，第25–54页。JMLR.org（2013）41. Xian, Y., Fu, Z., Muthukrishnan, S., de Melo, G., Zhang, Y.：用于可解释推荐的强化学习知识图谱推理。载于：第42届国际ACM SIGIR信息检索研究与开发会议论文集。第285–294页。SIGIR'19，计算机协会，纽约，美国（2019）42. Xian, Y., Fu, Z., Zhao, H., Ge, Y., Chen, X., Huang, Q., Geng, S., Qin, Z., de Melo, G., Muthukrishnan, S., Zhang, Y.：CAFE：用于可解释推荐的由粗到细的神经符号推理。载于：第29届ACM国际信息与知识管理会议论文集。第1645–1654页。CIKM '20，计算机协会，纽约，美国（2020）

43. Zhang, F., Yuan, N.J., Lian, D., Xie, X., Ma, W.Y.：面向推荐系统的协同知识库嵌入。载于：第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集。第353–362页。KDD '16，计算机协会，纽约，纽约州，美国（2016） 44. Zhang, Y., Chen, X.：可解释推荐：综述与新视角。信息检索基础与趋势 **14**(1)，1–101（2020） 45. Zhao, K., Wang, X., Zhang, Y., Zhao, L., Liu, Z., Xing, C., Xie, X.：利用示范进行知识图谱上的强化推荐推理，第239–248页。ACM，纽约，纽约州，美国（2020） 46. Zhao, W.X., He, G., Yang, K., Dou, H., Huang, J., Ouyang, S., Wen, J.R.：KB4Rec：用于连接知识库与推荐系统的数据集。Data Intelligence **1**(2)，121–136（04 2019） 47. Zhao, Y., Wang, X., Chen, J., Wang, Y., Tang, W., He, X., Xie, H.：面向推荐的知识图谱时间感知路径推理。ACM Trans. Inf. Syst.（2022年4月），刚刚录用 48. Zhou, T., Kuscsik, Z., Liu, J., Medo, M., Wakeling, J.R., Zhang, Y.C.：解决推荐系统表面上的多样性—准确性困境。美国国家科学院院刊 **107**，4511–4515（2010）