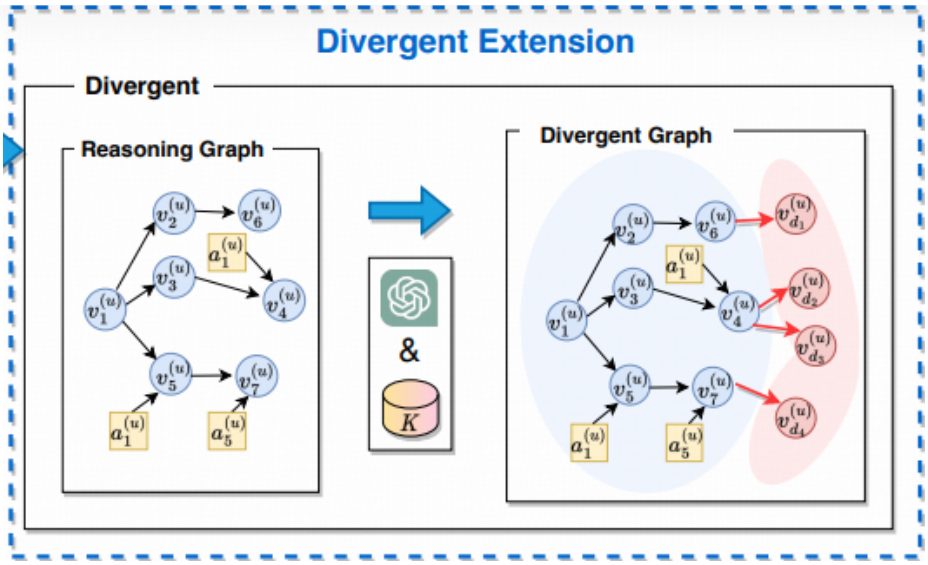


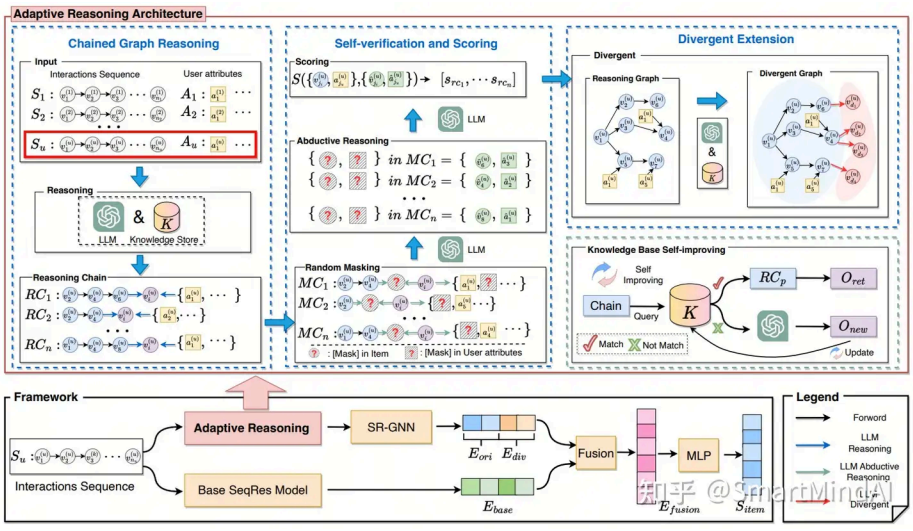
阿里2024：LLMRG-借助LLM的图模型助力推荐效果升级

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

Introduction

现有的推荐系统虽能个性化推荐，但受限于传统机器学习的逻辑解释，无法全面理解用户兴趣的动态演变。为弥补这一缺陷，研究者提出LLM推理图（LLMRG），利用大型语言模型✦进行深度因果逻辑推断，构建个性化的图结构。LLMRG克服了现有知识图谱✦推荐的局限，通过图神经网络处理生成的图，无需额外信息，就可以捕捉用户兴趣的深层次语义关系✦。实验证明，这种方法显著提高了推荐的逻辑性和可解释性，展示了在重构推荐系统方面的重要潜力。

LLMRG通过四个核心模块------连贯图推理、分歧扩展、自我验证与知识库自我改进，运用大型语言模型进行深度推理，形成个性化推荐图。此框架创新性地从用户行为历史中生成潜在的因果关系链。它不仅能验证现有推理，还能预测未来可能的兴趣。实验证明，LLMRG在提升推荐准确性和减少数据需求上表现出优越性，强调了语言模型在增强推荐系统逻辑性和解释性方面的价值。因此，LLMRG通过因果逻辑分析，不仅提供了即时反应的推荐，还具备前瞻性，能主动推荐新商品，实现了对用户需求更深入的理解和满足。



LLMRG

Problem

推荐系统的核心任务是在理解用户历史行为的基础上，预测用户未来可能感兴趣的商品。将用户交互历史视为一个序列，通过模型学习用户兴趣的动态变化。设用户集合为$\mathcal{U}$，物品集合为$\mathcal{V}$，每个用户u有交互序列

$$\mathcal{S}_u = [v_1, v_2, \ldots, v_{n_u}]$$

其中每个时间步的交互项$v_t^{(u)}$代表用户在特定时刻的选择n_u是交互项的总数。用户属性集$\mathcal{A}_u$包含个性化的属性信息，用于个性化建模，共有n_a个属性。

目标是基于用户历史和属性，预测用户在时间步$t = n_u + 1$会购买的物品，其概率分布为：

$$P(v_{n_u+1}^{(u)} | \mathcal{S}_u, \mathcal{A}_u)$$

这通常通过序列建模算法，如RNN、LSTM或Transformer，来实现，通过对用户行为模式的学习，生成对未知物品的预测概率分布。

$$p(v_{n_u+1}^{(u)} = v | \mathcal{S}_u, \mathcal{A}_u)$$

1. 利用大型语言模型（LLM），依据用户交互序列$\mathcal{S}_u$和属性集$\mathcal{A}_u$，创建个性化的推理图。利用LLM的因果推理和逻辑分析能力，揭示用户行为习惯和兴趣特点。
2. 这种图结构提供了用户兴趣的可解释模型，能够嵌入丰富的语义关系，帮助理解用户行为背后的深层次逻辑。
3. 我们设计了一种基于LLM能力自适应的推理架构，由四个关键组件组成，旨在优化信息提取和推荐过程：
4. 信息抽取：从LLM中提取与用户行为相关的隐含信息。
5. 关系建模：利用LLM的逻辑推断能力，建立用户与物品之间的潜在关系。
6. 特征融合：结合用户属性和交互信息，形成更全面的兴趣表示。
7. 预测模块：基于生成的图和用户属性，预测用户未来可能的交互行为。

通过这种方法，LLMRG不仅能提供精准的推荐，还能提高推荐的透明度，使用户更容易理解和接受推荐结果。

Adaptive Reasoning Architecture

LLMRG Framework

自适应推理模块和基础的顺序推荐模型。自适应推理模块处理用户交互序列$\mathcal{S}_u$和属性集$\mathcal{A}_u$，通过链式图推理、自我验证及评分，生成推理图G_{rea}和分歧图G_{div}。映射ϕ将这些信息转化为原始和分歧图的嵌入E_{ori}和E_{div}，分别由SR-GNN处理。基础推荐模型则生成E_{base}。最后，融合嵌入E_{fusion}通过函数ψ预测用户下一项的物品。

这种方法旨在捕捉用户动态兴趣，通过LLM的个性化推理图提高推荐的准确性。自适应推理模块创建个性化的推理图，而分歧扩展通过创新，不仅依赖用户行为反馈，还能主动发掘潜在兴趣，进行非线性推荐。自我验证和评分确保了推荐的逻辑连贯性和合理性。这种策略与基础的顺序推荐模型结合，利用各自优势，无需增加额外用户数据，实现了协同优化。

Experiments

Experiment Settings

实验使用三个数据集：Amazon Beauty、Amazon Clothing和MovieLens-1M，它们具有高稀疏性和短交互记录，分别来自Amazon和MovieLens。

Table 1: Statistics of the datasets after preprocessing.

Specs.	Beauty	Clothing	ML-1M
# Users	22,363	39,387	6,041
# Items	12,101	23,033	3,417
# Avg.Length	8.9	7.1	165.5
# Actions	198,502	278,677	999,611
Sparsity	99.93%	99.97%	95.16%

数据预处理后，关注商品类别和品牌作为特征。在MovieLens-1M，利用电影类型。我们以用户交互视为隐性反馈，去除重复并按时间排序。目的是全面评估方法性能。评估指标采用留一法，随机排除一个项目预测，考察模型对未观察项的预测能力，以HR@n（命中率）和NDCG@n（折现累计增益）来衡量，其中n取5和10。为了保证结果稳定，我们重复实验5次，使用不同的随机种子。

实验部分参照了三种基线方法：

(1) BERT4Rec，使用BERT双向Transformer进行通用序列建模；

(2) FDSA，结合属性信息，通过自注意力块进行属性感知推荐；

(3) CL4SRec和DuoRec，分别通过对比学习优化，前者采用数据增强辅助目标，后者创新性地结合监督和无监督学习。

这些方法均针对序列推荐，旨在展示我们的方法与现有技术的差异和优势。

实验结果与分析如下 1. 无LLM（纯推荐）：未使用LLM模块，直接运行基础推荐模型，结果显示性能显著下滑，证实LL的逻辑推理功能对推荐质量至关重要。

1. 非GPT基础：将LL模块替换为BERT，尽管有所波动，但仍显示出GPT4的优越性，表明不同基础模型对推荐有不同影响。
2. 单一任务学习：仅训练LL模块进行点击率预测，表现不佳，证实多任务学习策略（LLMRG）的优势。

Table 2: Performance comparison on three benchmark datasets, i.e., ML-1M, Amazon Beauty, and Amazon Clothing. We set the original models as baselines to compare with our proposed LLMRG model based on GPT3.5 or GPT4. The shaded area indicates the improved performance of our LLMRG model over the baselines across all three datasets. Higher is better.

Dataset	Metric	FDSA			BERT4Rec			CL4SRec			DuoRec		
		Original	GPT3.5	GPT4	Original	GPT3.5	GPT4	Original	GPT3.5	GPT4	Original	GPT3.5	GPT4
ML-1M	HR@5	0.0909	+ 20.70%	+ 25.79%	0.1124	+ 26.67%	+ 32.56%	0.1141	+ 19.98%	+ 21.02%	0.2011	+ 12.87%	+ 14.76%
	HR@10	0.1631	+ 17.93 %	+ 22.87%	0.1910	+ 13.52 %	+ 16.49 %	0.1866	+ 17.30 %	+ 19.31 %	0.2837	+ 14.10 %	+ 15.53 %
	NDCG@5	0.0599	+ 21.33%	+ 30.27 %	0.0713	+ 25.74 %	+ 32.82 %	0.0721	+ 14.97 %	+ 16.78 %	0.1265	+ 23.55 %	+ 26.01 %
	NDCG@10	0.0878	+ 21.78%	+ 28.25%	0.0980	+ 23.34%	+ 28.06%	0.1013	+ 17.67%	+ 20.42%	0.1663	+ 12.86%	+ 13.77%
Amazon Beauty	HR@5	0.0237	+ 13.89 %	+ 17.53 %	0.0201	+ 19.17 %	+ 23.22 %	0.0398	+ 11.15 %	+ 14.15 %	0.0552	+ 9.31 %	+ 11.93 %
	HR@10	0.0418	+ 15.02 %	+ 17.78 %	0.0413	+ 17.79 %	+ 22.14 %	0.0664	+ 10.22 %	+ 11.32 %	0.0839	+ 5.14 %	+ 6.61 %
	NDCG@5	0.0195	+ 16.20 %	+ 18.64 %	0.0192	+ 14.21 %	+ 17.63 %	0.0221	+ 8.45 %	+ 10.18 %	0.0350	+ 7.42 %	+ 9.24 %
	NDCG@10	0.0275	+ 14.78 %	+ 17.64 %	0.0263	+ 11.53 %	+ 14.76 %	0.0322	+ 8.17 %	+ 9.68 %	0.0447	+ 6.67 %	+ 7.95 %
Amazon Clothing	HR@5	0.0119	+ 20.67 %	+ 23.92 %	0.0128	+ 16.09 %	+ 19.10 %	0.0166	+ 7.90 %	+ 10.92 %	0.0190	+ 9.98 %	+ 11.40 %
	HR@10	0.0197	+ 14.45 %	+ 17.88 %	0.0202	+ 10.52 %	+ 13.72 %	0.0273	+ 11.21 %	+ 8.06 %	0.0311	+ 6.74 %	+ 9.19 %
	NDCG@5	0.0073	+ 8.16 %	+ 10.86 %	0.0081	+ 7.39 %	+ 10.39 %	0.0093	+ 6.02 %	+ 9.09 %	0.0118	+ 6.74 %	+ 9.19 %
	NDCG@10	0.0109	+ 6.01 %	+ 8.13 %	0.0113	+ 5.21 %	+ 5.94 %	0.0125	+ 4.32 %	+ 8.07 %	0.0155	+ 7.89 %	+ 9.29 %

实验揭示了LL在推荐中的核心价值，同时对比不同设置也验证了LLMRG通过整合大型语言模型和多任务学习实现性能提升的有效性。这强调了LLMRG模型设计的灵活性和适应性，能够满足推荐领域的复杂需求。

为了验证推理图的有效性，我们在ML-1M和Amazon Beauty数据集上对LLMRG模型进行了元组消融研究。首先，与DuoRec基线和不构建推理图的DuoRec+GPT3.5/4对比。序列图(DuoRec)仅处理交互序列，未进行推理，而LLMRG在处理交互的同时通过推理图进行深度理解。

Table 3: Ablation studies of our LLMRG model on two benchmark datasets, i.e., ML-1M and Amazon Beauty. We take the DuoRec as a baseline model to compare with the DuoRec with sequence graph and DuoRec with direct recommendation results via naive GPT3.5 or GPT4 without constructing a reasoning graph. The gray shaded area indicates our LLMRG's improved performance over the baseline across two datasets. The blue shaded area indicates the DuoRec with direct recommendation results via naive GPT3.5 or GPT4 without constructing a reasoning graph. Higher is better.

Method			ML-1M				Amazon Beauty			
			HR@5	HR@10	NDCG@5	NDCG@10	HR@5	HR@10	NDCG@5	NDCG@10
DuoRec			0.2011	0.2837	0.1265	0.1663	0.0552	0.0839	0.0350	0.0447
DuoRec	w/	seq graph	+ 6.36 %	+ 7.12 %	+ 12.25 %	+ 4.50 %	+ 3.26 %	+ 2.74 %	+ 3.71 %	+ 2.68 %
DuoRec+GPT3.5	w/o	rea graph	+ 0.94 %	+ 0.81 %	+ 0.55 %	+ 1.80 %	- 1.26 %	- 0.71 %	- 0.85 %	- 0.89 %
LLMRG (GPT3.5)	w/	rea graph	+ 12.87 %	+ 14.10 %	+ 23.55 %	+ 12.86 %	+ 9.31 %	+ 8.22 %	+ 9.71 %	+ 9.17 %
DuoRec+GPT4	w/o	rea graph	+ 3.28 %	+ 2.29 %	+ 3.95 %	+ 2.22 %	+ 0.72 %	+ 0.71 %	+ 0.86 %	+ 0.67 %
LLMR(										

结果显示，带推理图的LLMRG模型相对于DuoRec有显著提升，尽管GPT-4的强大力量使得DuoRec+GPT4在某些情况下表现更好，但总体上仍落后于LLMRG。 这些结果强调了推理图的重要性，它能通过清晰的推理过程提供精确且可解释的推荐，而单纯的GPT模型输出并不足以达到预期效果。LLMRG模型的成功表明，通过整合大型语言模型和推理图，能够更有效地利用用户信息和历史交互，从而优化推荐性能。元组消融实验进一步确认了推理图在利用大型语言模型于推荐系统中的关键作用，突出了其在提升推荐质量方面的价值。

我们进一步的消融实验深入研究了LLMRG各组件的效果。与DuoRec对比，我们发现不包括分歧扩展的LLMRG（无论是否使用GPT3.5或GPT4）性能提升有限，而去除自我验证模块时，性能下降，这揭示了GPT在未经验证时可能带来的噪声影响。这强调了分歧扩展和自我验证模块的重要性，它们能提升模型的推理能力，同时保持准确性。

Table 4: Ablation studies of our LLMRG model on two benchmark datasets, i.e., ML-1M and Amazon Beauty. We take the DuoRec as a baseline model to compare with the ablation models w/ or w/o divergent extension and self-verification modules based on GPT3.5 or GPT4. The shaded area indicates our ablation models' improved or decreased performance over the baseline across two datasets. Higher is better.

LLM	Method	ML-1M				Amazon Beauty			
		HR@5	HR@10	NDCG@5	NDCG@10	HR@5	HR@10	NDCG@5	NDCG@10
NA	DuoRec	0.2011	0.2837	0.1265	0.1663	0.0552	0.0839	0.0350	0.0447
GPT3.5	w/o div	+ 5.12 %	+ 3.87 %	+ 8.30 %	+ 4.75 %	+ 3.62 %	+ 2.86 %	+ 4.57 %	+ 3.80 %
	w/o ver	- 4.72 %	- 3.94 %	- 10.90 %	- 4.14 %	- 2.17 %	- 1.43 %	- 2.57 %	- 2.46 %
	w/ div & ver	+ 12.87 %	+ 14.10 %	+ 23.55 %	+ 12.86 %	+ 9.31 %	+ 5.14 %	+ 7.42 %	+ 6.67 %
GPT4	w/o div	+ 7.06 %	+ 4.68 %	+ 13.35 %	+ 8.11 %	+ 4.89 %	+ 2.15 %	+ 4.28 %	+ 5.81 %
	w/o ver	+ 5.86 %	+ 2.36 %	+ 5.77 %	+ 3.72 %	+ 1.26 %	+ 0.72 %	- 0.57 %	+ 1.34 %
	w/ div & ver	+ 14.76 %	+ 15.53 %	+ 26.01 %	+ 13.77 %	+ 11.93 %	+ 6.61 %	+ 9.24 %	+ 7.95 %

原文《Enhancing Recommender Systems with Large Language Model Reasoning Graphs》

编辑于 2024-05-24 15:53 · IP 属地北京

推荐系统　LLM（大型语言模型）　图模型

理性发言，友善互动

还没有评论，发表第一个评论吧

推荐阅读

2023关于LLM的总结

1、如果只做单轮的场景，完全可以只做文本向量的检索，要不要放LLM不影响单轮。如果想兼顾其它通用知识的回答，可以做RAG。
2、如果要做多轮的场景，可以判断一下这个场景，是不是

第十章 LLM评估指标

LLM 评估指标：LLM 评估所需的一切自2022年，ChatGPT发布之后，大语言模型（Large Language Model，简称LLM）掀起了一波狂潮。 本文将深入探讨：LLM 评估指

LLM大模型创业三大军规

上周，一家初创公司未能围绕和 RAG 开展业务，尽管他们第一份 B2B 大型合同。 以下以及如何避免这种情况： 创始了一篇博客解释了为什么他不

我在这里...