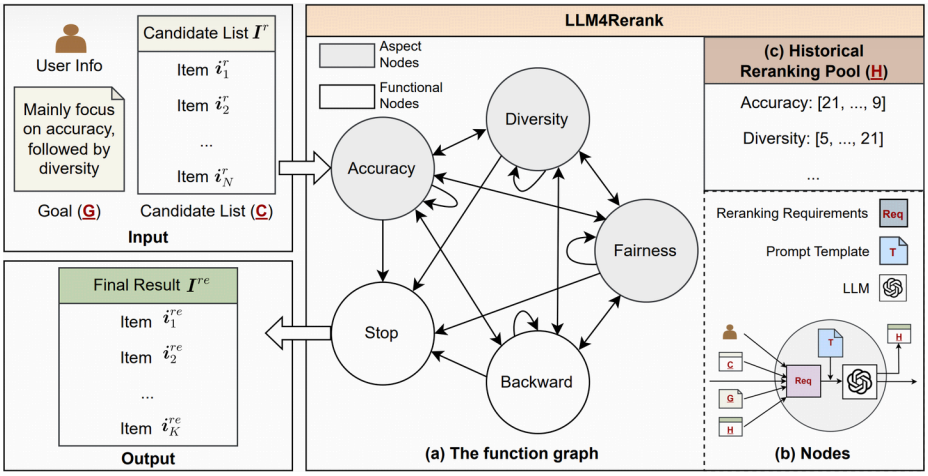


华为2024-LLM在推荐系统重排序中的应用

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

重排是推荐系统领域的基本技术，其重要性在于它能够细化并增强排序模型生成的结果，最终为用户提供最相关和个性化的推荐。目前，现有的重排模型主要集中在提高推荐的准确性方面。虽然准确性很重要，但在实际应用中，推荐列表的多样性和平等性也同样重要。多样性和公平性确保了用户能够接触到各种各样的内容，同时保证了不同类别或店铺的平等展示和曝光。尽管有一些研究尝试将其中一个与准确性方面结合进行建模，但如何更好地同时考虑和平衡更多方面的问题仍然存在。

目前，随着大型语言模型（LLMs）的快速发展，大量研究已经致力于评估LLMs在不同情境下的能力。先前的研究表明，零样本LLMs可能在信息抽取和推荐等任务上的专业性不如专门模型。这一限制通常归咎于它们的有限令牌数量和处理包含数千个内容的大量上下文的困难。尽管存在这些挑战，但这些研究已经证明，LLMs在重新排序任务中可以与监督基准相匹敌甚至超越，这归功于它们在涉及有限内容数量的紧凑上下文中的强大语义理解能力。因此，将LLMs应用于推荐的重新排序阶段，作为融合不同方面的一种关键策略，通过增强语义理解来提升效果。 然而，使用LLMs构建重新排序框架时，会遇到几个重大挑战。第一个挑战是确保框架在组织良好和灵活的方式下具有可扩展性，既能满足当前方面的需求，也能适应未来可能出现的方面。第二个挑战是设计一种机制，能够根据特定的推荐设置或用户偏好自动整合多样化的方面需求，最终实现真正的个性化。

为了应对这些挑战，我们提出了一种名为 LLM4Rerank 的重排序框架，它利用零样本LLM的能力以实现更精确的重排序。具体来说，LLM4Rerank 将重排序的各种方面要求表示为不同的节点，允许框架以链式思考（链式思考）的方式自动包含这些节点。

这种方法的优势在于两点：它确保了可扩展性，允许无缝地添加新节点以应对出现的新方面要求。为了展示这一点，除了准确性方面之外，本文还加入了多样性和公平性作为 LLM4Rerank 模型的方面，进一步丰富了其功能和潜力。

此外，LLM4Rerank 中使用的LLM可以自动确定下一个节点的考虑，这由当前重排序历史和用户或部署者提供的称为"目标"的额外句子输入引导，代表了正在进行的重排序过程的整体焦点和目标。这个动态过程使 LLM4Rerank 能够在重排序过程中实现增强的个性化。

Framework

Problem Formulation

知乎

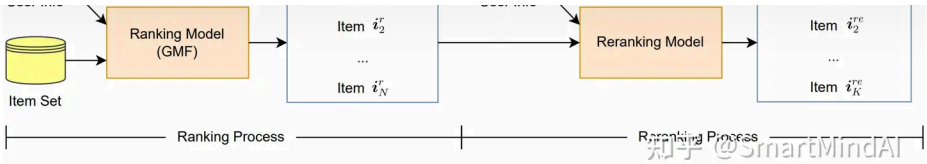


Figure 1: Ranking and reranking process in recommendations.

重排任务在推荐系统中扮演着至关重要的角色。如图所示，将用户集合表示为$\boldsymbol{U}$，将可推荐的内容集合表示为$\boldsymbol{I}$。为了清晰起见，本文将每个用户和内容表示为一个特征向量⁺（分别为$\boldsymbol{u}$和$\boldsymbol{i}$）。最初，一个排序模型为每个用户生成一个候选内容列表

$$\boldsymbol{I^r} = \{\boldsymbol{i_1^r}, \ldots, \boldsymbol{i_n^r}, \ldots, \boldsymbol{i_N^r}\}$$

其中包含$\boldsymbol{N}$个内容。为了提升推荐性能，会对初始列表$\boldsymbol{I^r}$中的内容之间的关系进行分析。这一分析旨在从初始列表生成一个精炼的列表，包含$\boldsymbol{K}$个内容，表示为

$$\boldsymbol{I^{re}} = \{\boldsymbol{i_1^{re}}, \ldots, \boldsymbol{i_k^{re}}, \ldots, \boldsymbol{i_K^{re}}\}（\boldsymbol{K < N}）$$

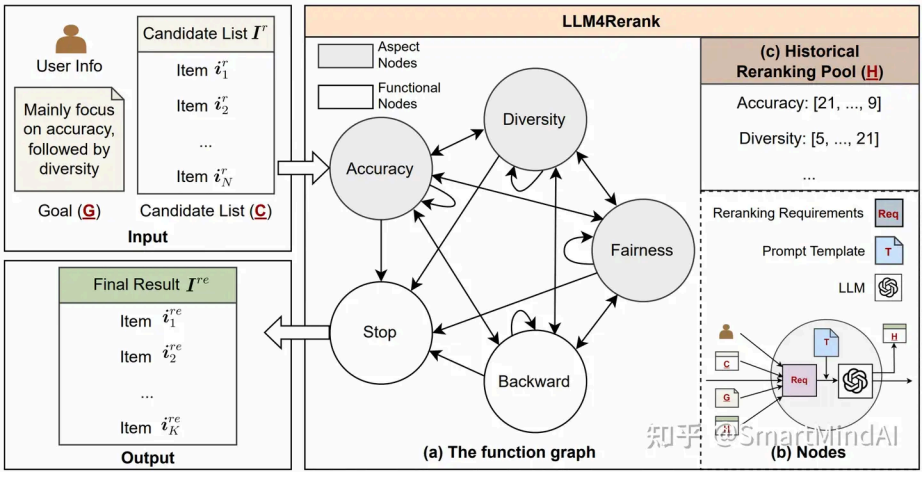
重排模型的目标是通过优化定义的目标函数⁺来增强用户与内容的相关性。

$$\boldsymbol{I^{re}} = \underset{\boldsymbol{i \in I^r}}{\boldsymbol{TopK}}\, \boldsymbol{R(u, i)}$$

其中 $\boldsymbol{R(u, i)}$ 是 评分函数，用于衡量用户$\boldsymbol{u}$对内容$\boldsymbol{i}$的相关性，考虑了准确性、多样性和公平性等多方面因素。

通过这一优化过程，重排序模型从候选列表$\boldsymbol{I^r}$中，根据个人用户的偏好定制，选择出最优推荐列表$\boldsymbol{I^{re}}$，从而显著提高了推荐的质量。 在本研究中，为了促进不同重排序基准之间的公平比较，遵循先前的研究，采用通用全局排序模型GMF。这种方法确保了所有重排序模型都在相同的候选内容列表上操作，便于进行统一的评估和比较。

LLM4Rerank Overview



如图所示, LLM4Rerank 框架接收三种类型的输入：用户信息（用户信息），包括如性别和年龄等特性；候选内容列表$\boldsymbol{I^r}$；以及一个称为"目标"的句子，概述了重新排序的优先方面。每个节点代表一个潜在的重新排序步骤，通过LLM生成重新排序列表，考虑特定的方面相关或功能需求，如图（b）所示。功能图形中的每条边都表示节点间的潜在过渡路径，确保所有节点之间的连通性，除了"停止"节点。为了展示，我们不仅包含了一个"准确性"节点，还包含了"多样性"和"公平性"方面的节点，以及两个功能性节点："后退"和"停止"，用于实际功能。节点结构精心设计，允许LLM依次评估不同的节点，从而优化重新排序结果，全面满足多个方面的需要。此外，为了防止遗忘并增强LLM对方面组合的评估，我们使用了一个历史重新排序池（图c）。这个池子记录了每个节点的连续结果，作为后续节点重新排序的辅助信息⁺。最终，当到达"停止"节点时，重新排序过程完成。这一阶段的输出表示为$\boldsymbol{I^{re}}$，即历史重新排序池中的最新重新排序结果。

Nodes Construction

主要挑战：首先，当引入额外的需求时，定制节点结构以确保其可扩展性；其次，使LLM能够自动选择其后续重新排序步骤。 为了解决这些挑战，我们引入了一个通用的节点结构。它包含一个重新排序步骤，与一个辅助指标相结合，该指标表示由下一个节点名称确定的后续步骤的方向。这种配置允许LLM自动通过当前可用的信息做出决策。

具体地，我们引入了一个通用的节点结构，如图（b）所示。这个通用节点代表了在LLM的考虑下执行单个重新排序步骤。通用节点的输入包括用户信息的语义表示、候选内容、定义整个重新排序过程个性化焦点的"目标"句子，以及如果可用，则包含整个历史重新排序池。此节点的输出有两个：当前节点的即时重新排序结果（以内容ID列表表示），并整合到历史重新排序池中，作为后续步骤的参考。此外，还会生成一个指示符（即本文中的下一个节点名称），用于指定后续节点的重新排序，从而实现自动步骤化的过程。

Aspect Nodes

为了帮助大型语言模型执行针对不同方面需求定制的重新排序任务，我们采用了一种基于提示的模板方法在所提出的通用节点结构中。这种方法允许在重新排序过程中实例化专门用于评估不同方面的特定节点。因此，每个节点都设计为系统性地处理这些关键方面之一，以确保重新排序结果反映出对各方面均衡考虑。

在本研究中，为了展示大型语言模型的可扩展性，我们实现了三个方面节点，用于重新排序："准确性"，"多样性"，和"公平性"。

- 准确性节点：该节点旨在在重新排序阶段满足最终推荐列表的准确性标准。因此，提示模板被精心设计以强调用户与内容的相关性。图展示了在节点中使用的简单模板实例。此外，鉴于推荐准确性的重要性------这是推荐系统不可或缺的基本方面------准确性节点已被确立为命名框架中的起始点。因此，每一次重新排序过程都从准确性节点开始，确保从一开始就专注于精确性的基础聚焦。
- 多样性节点：此节点专门设计用于评估最终推荐列表在重新排序阶段的多样性标准。在本研究中，我们通过评估内容属性的具体方面在最终列表中的变化程度，来评估重新排序结果的多样性方面。我们采用了α-NDCG指标进行这一评估。一个用于多样性节点的示例模板在图中展示，以直观地呈现这一过程。
- 公平节点：在重新排序阶段，此节点用于确保最终推荐列表中的公平性目标得以实现。在我们的研究中，公平性被量化为根据一个独特的特征分离的两个样本组之间的内容得分差异的均值。由于LLM本质上生成的是重新排序列表而不是数值得分，我们为最终推荐列表中的内容分配了从0到1的线性得分范围。这些得分随后用于计算所需的平均绝对偏差（MAD）以进行公平评估。

Functional Nodes

近期的研究已经证明了反思在优化LLMs的输出方面的有效性。为了增强，我们开发了两个功能节点，专门用于在重新排序序列中促进反思和终止操作。

- 反向节点：此节点赋予LLM选择性忽略的能力，在评估先前尝试时认为不理想的重排结果。从历史重排队列中移除最新的重排结果，并按照LLM的输出指令，继续执行到下一个节点。一个说明性的模板示例在图中给出。
- 停止节点：此节点管理输出序列的终止。当名称将此节点指定为输入步骤时，它标志着完整的重新排序过程结束。随后，此节点从历史重新排序池中提取出最新的重新排序结果，作为最终的重新排序结果呈现。请注意，此节点仅表示重新排序过程的结束，无需访问LLM。因此，这个节点不需要提示模板。

Automatic Reranking Process

为了利用语言模型根据多样化的方面要求进行重新排序，我们设计了独特的节点，每个节点都针对特定的方面标准。然而，为每项重新排序任务预定义从一个节点到另一个节点的路径既不高效，也难以实现。因此，为了适应独特的用户偏好并显著提高个性化，我们开发了一个自动重新排序过程，主要包含以下三个子过程：

1. 需求识别过程：此过程用于识别用户的具体需求和偏好，以便为每个重新排序任务提供个性化的节点路径。

知乎

3. 结果整合与个性化调整过程：此过程整合重新排序的结果，并根据用户的具体反馈进行个性化调整，以提供最符合用户预期的重新排序结果。

Algorithm 1: The whole automatic reranking process of LLM4Rerank

Input: User information $\boldsymbol{u}$, Candidate item list $\boldsymbol{I}^r$, the reranking focus $Goal$, Maximum node count MC
Output: Final reranking result $\boldsymbol{I}^{re}$
Note: $Function(a)(b)$ represents the execution of functions in node a with input b.

Initialize current node name $CN = Accuracy$; Current reranking result $CR = None$; Node count $NC = 0$; Historical reranking pool $Pool = []$.
while $CN \neq Stop$ **do**
  $CN, CR = Function(CN)(\boldsymbol{u}, \boldsymbol{I}^r, Goal, Pool)$
  $Pool.append(CR)$
  $NC += 1$
  if $NC \geq MC$ **then**
    $CN = Stop$
  end if
end while
return $Pool[-1]$

Experimental Setup

Dataset

我们使用三个广泛采用和认可的公共数据集进行实验：ML-1M，KuaiRand（KuaiRand-Pure），以及Douban-Movie 。对于每个数据集，我们采用留一法，将数据分为训练集、验证集和测试集[+]。

Dataset	Interactions	Users	Items
ML-1M	1,000,209	6,040	3,883
KuaiRand	102,433	10,494	7,583
Douban-Movie	759,652	2,606	34,893

Overall Performance (RQ1)

Table 2: Overall performance comparison. Symbols "-A/D/F" represent different focuses "Accuracy/Diversity/Fairness" when setting "Goal" in LLM4Rerank. The default LLM backbone is Llama-2-13B. ↑: higher is better; ↓: lower is better.

Model	ML-1M				KuaiRand				Douban-Movie			
	HR ↑	NDCG ↑	α-NDCG ↑	MAD ↓	HR ↑	NDCG ↑	α-NDCG ↑	MAD ↓	HR ↑	NDCG ↑	α-NDCG ↑	MAD ↓
GMF	0.4156	0.1853	0.1005	0.0613	0.4417	0.2314	0.1627	0.1588	0.5723	0.3150	0.2516	0.4006
DLCM	0.5781	0.2354	0.1378	0.0549	0.6893	0.3080	0.1767	0.1026	0.6827	0.4102	0.3581	0.2619
PRM	0.6986	0.3246	0.1653	0.0436	0.8083	0.3904	0.1869	0.1032	0.6979	0.4167	0.3477	0.2509
MMR	0.4675	0.2588	0.2104	0.0265	0.4928	0.2606	0.1877	0.1569	0.6538	0.3873	0.3744	0.2539
FastDPP	0.4719	0.2561	0.1942	0.0263	0.5728	0.2913	0.1882	0.0660	0.6635	0.4038	0.3818	0.2820
FairRec	0.4805	0.2007	0.1243	0.0199	0.6083	0.2761	0.1540	0.0318	0.6771	0.4021	0.3119	0.1752
RankGPT	0.5584	0.2587	0.1799	0.0564	0.6583	0.2910	0.1557	0.1256	0.6635	0.3967	0.3448	0.2472
GoT	0.5730	0.2714	0.1942	0.0486	0.7184	0.3198	0.1788	0.1211	0.6827	0.4135	0.3592	0.2195
LLM4Rerank-A	**0.7031***	**0.3320***	0.2294	0.0434	**0.8252***	**0.4229***	0.2032	0.1969	**0.7041***	**0.4301***	0.3806	0.2446
LLM4Rerank-D	0.6875	0.3292	**0.2407***	0.0571	0.8058	0.4143	**0.2223***	0.0969	0.6791	0.4019	**0.3837***	0.2757
LLM4Rerank-F	0.5584	0.2328	0.1411	**0.0193***	0.7282	0.3276	0.1825	**0.0271***	0.6698	0.3917	0.2570	**0.1656***
LLM4Rerank-ADF	0.6364	0.3058	0.2051	0.0250	0.8000	0.4117	0.2163	0.0530	0.6877	0.4105	0.3664	0.1975

比较结果显示：

知乎

- MMR 和 FastDPP 在通过α-NDCG 指标量化的增强方法中展示了增强多样性的有效性。这两种模型在内容相似性和整体列表多样性上表现出色，强调了这两点，从而为用户优化了重新排序的列表。
- 模型FairRec在使用MAD指标衡量的情况下，展现出在促进公平性方面的卓越能力。通过融合分解对抗学习与正交性+正则化技术，模型FairRec确保不同用户群体获得更公平的推荐，实现了跨用户组+的更均衡的建议。
- RankGPT展现出卓越的性能，强调了零样本LLMs在重新排序任务中的能力。相反，GoT通过采用思考链策略，促进了对多个方面的顺序分析，取得了更优秀的成果。
- 通过个性化"目标"设置和自动重新排序过程，LLM4Rerank 显著超越了基准线，全面验证了其有效性。精妙地结合了语言模型的LLM4Rerank，展示了其在重新排序方面的灵活性和广泛适应性。虽然LLM4Rerank-ADF 在任何单一方面可能不处于领先地位，但其在所有维度上的均衡表现证实了将LLMs与自动重新排序框架结合的优势。这种方法通过语义理解有效地调和了不同方面的需求，提供了在准确度、多样性和公平性方面优化的结果。

原文《LLM-enhanced Reranking in Recommender Systems》

发布于 2024-08-06 11:48 · IP 属地北京

推荐系统 LLM 排序

推荐阅读

考美国Bar?深度解析LLM发展前景

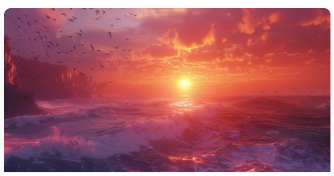
近些年，赴美国读LLM的持续升温，但许多人对LLM的发展前途仍不十分掌握。 文中将对LLM普遍的三条关键发展方向：1）留美考刑事辩护律师营业执照（考BAR）；2）归国；3）留美转法律学学士（JD...

宏景涉外律师

llm的信息抽取、理解、推理的对比选型（一）

背景 因实际的业务需要，计划引入llm，发挥其实体识别，信息抽取（比如个人信息）的作用；同时兼顾意图识别能力，首先基于公开的大模型公开评测结果 https://opencompass.org.cn/，...

不亦乐乎



产品角度的LLM落地思考

晓风暮笛

我为什么不看好LLM ——去一年实习经历有感

简介笔者为了寻求科研界的不验，试图在工业界寻求一些经在大三一整年做过差不多9个习，分别在三家独角兽公司担法工程师，参与模型算法和/Infra工程，从中吸取了大...

JerryYin777