LLM4SBR 轻量框架：实现对话式推荐中 LLM 的整合与工业级部署

原创 方方 方方的算法花园 2024年11月12日 08:46 北京

0 论文概况

1. 论文名称：
LLM4SBR: A Lightweight and Effective Framework forIntegrating Large Language Models inSession-based Recommendation
《LLM4SBR：一种将LLM整合到基于会话的推荐中的轻量级有效框架》
2. 论文链接：https://arxiv.org/pdf/2402.13840
3. 论文作者所在机构：重庆大学、腾讯、清华大学
4. 一句话概括：该论文提出了用于基于会话推荐（SBR）的轻量级有效框架 LLM4SBR，通过将大语言模型（LLM）与 SBR 模型结合，分两步处理会话数据，解决了传统 SBR 模型缺乏语义信息及 LLM 应用于 SBR 的诸多问题，实验证明其能显著提升传统 SBR 模型性能且适合工业部署。

1 挑战

1. 传统 SBR 模型的局限性
（1）语义信息缺失：传统 SBR 研究基于 ID - 行为推荐范式，常使用 one - hot 编码表示项目，虽能高效建模协同信息，但忽视了交互行为中的语义信息，如项目名称、价格等。在 SBR 中，序列长度短且数据稀疏，仅建模稀疏行为信息难以理解用户真实意图。例如，用户点击序列为 “iPhone 15”“跑步鞋”“iPhone 14”“牛奶”“裙子”，仅从行为建模可能误判 “牛奶” 和 “裙子” 为关键兴趣，而语义信息可分析出用户可能对苹果产品系列更感兴趣。

2. LLM 与 SBR 结合的困难
（1）LLM 幻觉问题：SBR 序列长度通常较短，且无法获取用户个人信息，导致 LLM 可利用信息有限，容易出现无法生成有效答案或生成超出项目集的虚假项目的情况。
（2）“repeater” 问题：SBR 数据通过序列分割增强，会产生包含大量相似会话的数据集，微调 LLM 时可能使模型过度重复输入文本或生成重复句子。
（3）资源消耗问题：LLM 计算复杂，占用大量 GPU 内存且推理时间长，而推荐任务追求实时性，基于 LLM 的 RS 模型难以在工业实践中实现。

2 论文贡献点

1. 提出首个LLM增强框架：
率先提出适用于SBR的LLM增强框架LLM4SBR，将LLM推理和SBR模型训练分为两个阶段，提前将LLM推理结果保存于外部文件，确保训练时GPU使用和训练时间仅取决于SBR模型，有效解决了传统SBR模型缺乏语义信息理解能力以及LLM与SBR结合时面临的高成本、易出现幻觉等问题。

2. 设计意图定位模块：提出意图定位模块，通过计算推理结果与项目集文本嵌入的余弦相似度，筛选出最相似的实际项目来修正LLM推理结果，消除幻觉并增强语义。同时，对不同视

角的embedding进行对齐和统一，实现了更细粒度的模态对齐，促进了交互ID信息和文本信息的有效融合。

3. 显著提升模型性能：在两个真实数据集上的实验表明，LLM4SBR框架可应用于大多数当前SBR模型，能大幅提高模型性能，在预测准确性、模型可解释性和候选多样性等方面均有显著提升，且框架具有轻量级和高效的特点，适用于工业部署场景，为SBR研究和工业应用提供了新的有效方法。

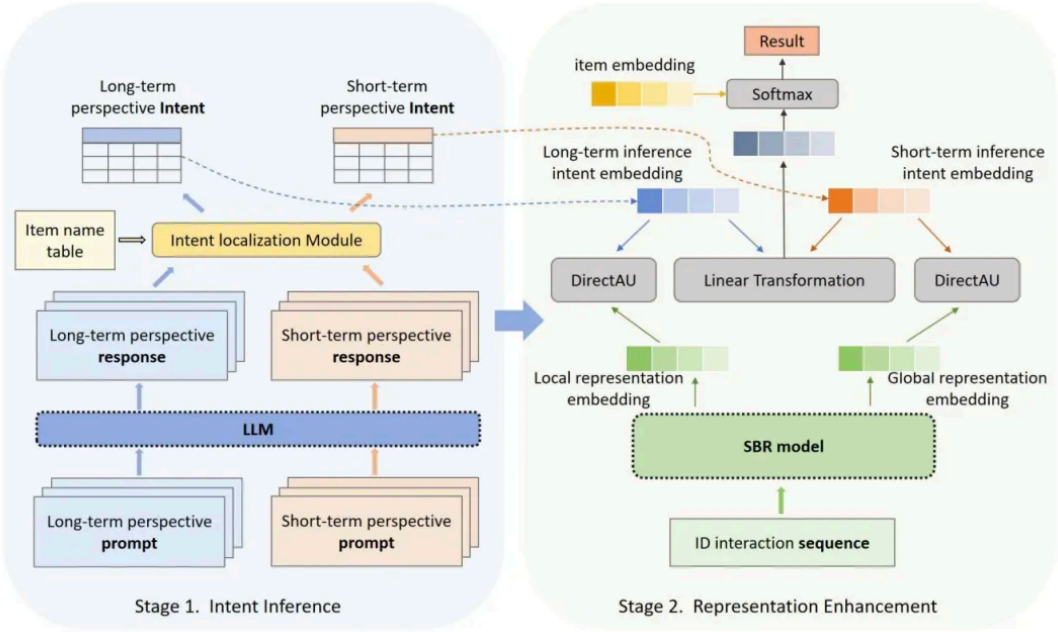
3 LLM4SBR 框架

问题定义

SBR 的目标是预测匿名用户当前会话历史中可能的下一个交互项目。用数学符号定义了数据集、会话、项目集等相关概念，明确了建模目标是基于会话的历史行为记录预测下一个点击项目。

LLM4SBR整体架构

LLM4SBR 框架分为**意图推断和表示增强**两个阶段，先利用 LLM 对会话文本数据进行推理，再用 SBR 模型处理行为数据，最后融合多视角和多模态的会话表示用于推荐。



意图推断

（1）Prompt 设计：为更准确推断意图，引入基于 SBR 常用行为建模视角（长、短期）的限定词设计 prompt，将文本推理任务分解为更细粒度的子任务，提高 LLM 推理能力利用效率。prompt 包含背景、项目名称序列和任务提示三部分，还加入了项目 ID 信息，示例了长、短期 prompt 模板及 LLM 推理输出要求。

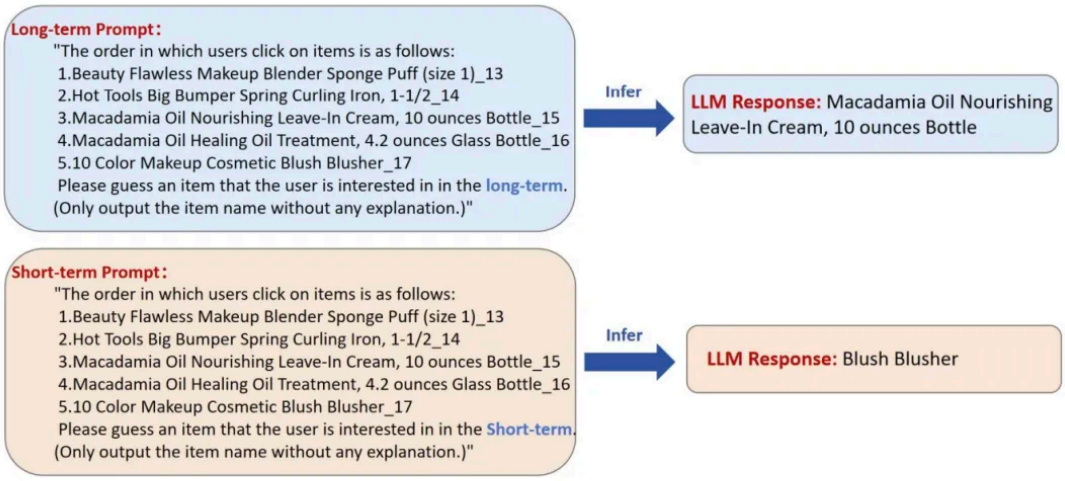


Figure 2: Illustration of the design of prompts.

（2）LLM 推理：选择 Qwen 7B 模型作为推理模型，以问答形式输入 prompt 获取推理结果，通过在 prompt 中标记特定输出要求来规范 LLM 回答，同时强调模型可替换，参数更多、推理能力更强的 LLM 能产生更准确结果。

（3）意图定位：为解决 LLM 幻觉和语义增强问题设计此模块。受 RAG 检索模型启发，先将推理结果和项目集文本编码为embedding形式，计算两者余弦相似度，选取相似度最高的 Top - f 个实际项目，通过加权求和得到优化后的 LLM 推理结果，最后将调整后的结果embedding存储于外部文件，减少后续计算时间。

表示增强阶段

（1）SBR 建模：使用 SBR 模型建模会话序列中的交互信息和学习用户行为偏好，以经典的 SR - GNN 模型为例，它将会话数据构建为会话图，用 GGNN 学习节点特征，以最后点击项目为局部嵌入，聚合节点信息并通过软注意力机制表示全局偏好，同时指出框架中的 SBR 模型可替换，并在实验部分测试了替换后的性能。

（2）文本嵌入解析：读取意图推断阶段存储的推理结果embedding，通过特定函数转换为张量形式并进行维度对齐，为后续融合做准备。

（3）表示对齐和融合：由于 SBR 模型和 LLM 推理的embedding不在同一空间，使用 DirectAU 方法分别计算不同视角下推理表示与会话表示的对齐损失，以及各自内部的均匀损失，以更好地整合embedding。然后将不同视角和模态的会话表示融合为最终用于预测的会话表示，通过线性层压缩相关嵌入到同一空间。

（4）预测和优化：通过计算会话表示与项目表示的得分，经 softmax 函数得到预测值，定义 SBR 任务的损失函数为预测值与真实值的交叉熵，最终的联合学习损失函数由推荐损失和辅助任务（对齐和均匀）损失组成，通过控制参数 τ 调整辅助任务比例。

4 实验与结论

文中进行了性能实验，主要对比了在不同的 Top-K下，SBR 模型和应用 LLM 框架的 SBR 模型的性能。具体内容如下：

实验设置

（1）数据集：选择了 Beauty 和 MovieLens-1M（ML-1M）数据集。由于没有提供交互 ID 序列和项目名称信息的常用 SBR 数据集，所以选用了这两个格式较为接近的数据集。其中，Beauty 数据集包含用户对各种美容产品的评价和评分，将单个用户的所有评分序列视为一个会话序列，并使用常用的序列分割方法进行数据增强；ML-1M 数据集则包含了超过 6000 名用户对超过 4000 部电影的 100 多万条评分，以 10 分钟为间隔将同一用户的电影评分数据划分为多个会话序列。同时，按照相关研究，去除了长度为 1 的会话和在所有会话中出现次数少于 5 次的项目。

（2）评估指标：选用了 SBR 任务中最常用的指标，即 Precision（P）@K和 Mean reciprocal rank（MRR）@K，并且将候选集 @K的长度设置为 5、10 和 20，以便进行对比。

（3）参数设置：所有实验均在 NVIDIA A100 GPU 上进行，统一使用 Adam 优化器，学习率为 0.001，每三个 epoch 衰减 0.1，L2 惩罚设置为 10^−5。实验中涉及的 SBR 模型的批量大小为 100，维度大小为 100，超参数 τ 设置为 0.1，意图定位模块中的超参数 f 初始设置为 5，后续超参数实验将讨论其最优值。其他参数则按照相应论文中的最优设置进行。

选用的 SBR 模型

选择了四个经典的 SBR 模型来替换框架中的 SBR 模型以验证框架的有效性，具体如下：

- **SR-GNN：**第一个将会话数据构建为会话图的模型，利用 GGNN 来捕获项目间复杂的转换关系。
- **TAGNN：**在 SR-GNN 的基础上添加了目标敏感的注意力机制。
- **GCE-GNN：**分别构建会话图和全局图，从项目级别和会话级别学习相关信息。
- **S²-DHCN：**使用超图卷积来学习项目序列中的高阶关系，并使用自监督学习来缓解超图的数据稀疏问题。

实验结论

1. LLM4SBR 显著提升了模型的性能。通过 LLM 框架增强的所有模型均表现出性能提升，这表明从 LLM 推理得出的文本表示包含丰富且有价值的信息，能极大地帮助 SBR 模型理解会话数据的潜在意图。

2. LLM4SBR 对较小的K值有更大的提升。例如，LLM4SBR（TAGNN）使两个数据集的 P@5 指标分别提高了 27.28% 和 107.5%，这是由于 LLM4SBR 在意图定位阶段利用了 f 个相似语义项目实现语义增强，从而使预测候选集中排名靠前的项目更加准确。

3. LLM4SBR 可以弥补因交互信息缺乏导致的建模不佳。GCE-GNN 通过同时构建全局图和会话图来捕获有效信息，但由于其模型计算复杂，在数据量有限的情况下难以

学习到有效的会话表示。而 LLM4SBR（GCE-GNN）表现出最大的改进，特别是在 ML-1M 数据集上，P@5、P@10 和 P@20 分别增加了 37.59%、96.2% 和 128.54%，这归因于从 LLM 推理获得的有效文本信息，弥补了 GCE-GNN 会话建模中的信息稀缺，使其能够获得更好的性能。

4.与框架集成后，S²-DHCN 和 GCE-GNN 在一些指标（P@20 和 MRR@20）上的性能略有下降。作者认为当原始 SBR 模型已经有效地对数据进行建模时，通过意图定位模块增强推理信息可能会引入噪声，但与改进幅度相比，这种下降非常轻微。此外，由于可以通过调整意图定位模块中的超参数 f 来有效控制噪声问题，因此负面影响几乎可以忽略不计。

END

LLM与推荐 15 LLM论文阅读 13

LLM与推荐 · 目录

上一篇 亚马逊COSMO：LLM构建高质量电商知识图谱

下一篇 RecRanker：指令调优LLM用于 top-k 推荐排序