LLMs 推荐发展综述-工业部署（加速|冷启动|动态更新|需求定制）

原创 方方 方方的算法花园 2024年11月11日 09:38 北京

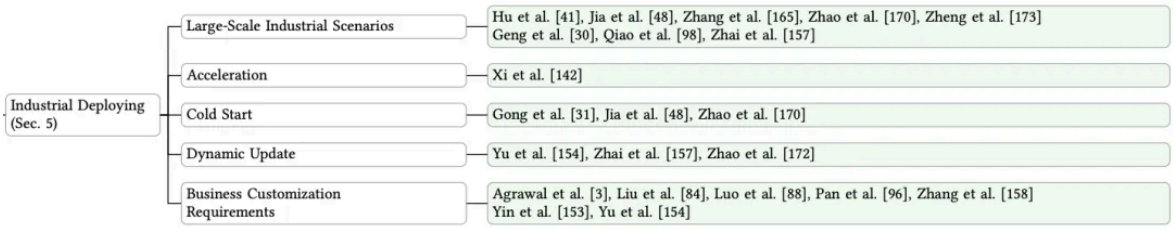
点击蓝字 关注我们

写在前面

吉林大学、悉尼科技大学、香港理工大学在2024年10月联合发表了一篇论文《Towards Next-Generation LLM-based Recommender Systems: A Survey and Beyond》，论文里探讨了基于大语言模型（LLMs）的推荐系统的发展，包括其在表示与理解、设计与应用、工业部署等方面的应用，以及面临的挑战和机遇。（论文链接：https://arxiv.org/pdf/2410.19744）

本文主要介绍论文中关于 Industrial Deploying 部分的研究进展，关注LLMs 推荐在大规模工业场景中的方法、加速、冷启动、实施动态更新以及满足各种业务定制需求，更好地理解在现实世界工业应用中部署基于大型语言模型的推荐系统的当前进展和实际考虑。

其他部分内容请参考此系列其他文章。



大规模工业场景

在大规模工业应用中，部署基于LLMs的推荐系统由于数据量庞大以及用户和项目动态不断变化，带来了巨大的复杂性。这些环境需要高效处理广泛的特征空间，同时满足多样化和不断发展的业务需求。这些系统的规模凸显了在计算效率和保持高质量推荐之间取得平衡的必要性。传统的推荐pipelines由更小、更专业的模型组成，在成本和更新方面更容易管理。相比之下，LLMs的资源需求——特别是在训练和推理方面——使它们的大规模部署更具挑战性。**LLMs并非完全取代传统模型，而是越来越多地被整合为组件，以提高整体性能。**

[170]Breaking the barrier: Utilizing large language modelsfor industrial recommendation systems through an inferential knowledge graph. arXiv preprint arXiv:2402.13750,2024.

提出的 LLM-KERec 方法将传统模型高效的协同信号处理能力与LLM和互补图相结合。这种方法不仅减少了传统模型推荐结果的同质性，还提高了整体点击率和转化率，使大型语言模型能够在工业场景中大规模应用。

[173]A largelanguage model enhanced sequential recommender for joint video and comment recommendation. arXiv preprintarXiv:2403.13574, 2024.
指出，LLM的计算成本使其难以在大规模工业环境中有效地部署为在线推荐系统。因此，这项工作提出仅在训练阶段使用大型语言模型，以补充和增强我们推荐主干的语义能力。

[165]Recgpt: Generative personalizedprompts for sequential recommendation via chatgpt training paradigm. arXiv preprint arXiv:2404.08675, 2024
为了解决大型模型对于在线推荐服务过于沉重的问题，探索了如何灵活有效地将 ChatGPT 整合到推荐系统中。他们放弃了自然语言形式，而是使用 ChatGPT 的模型结构和训练范式进行项目序列预测。

[48] Knowledge adaptation from large language model to recommendation for practical industrial application. arXivpreprint arXiv:2405.03988, 2024.
指出，在工业场景中，基于用户交互历史进行推理的大型语言模型或对其进行微调是不切实际的。为了解决处理大量用户交互历史所带来的计算挑战，CEG 模块使用预训练的大型语言模型作为项目编码器，而不是用户偏好编码器，从而降低了计算开销。

[41]Enhancing sequential recommendation via llm-based semantic embedding learning. In Companion Proceedings of theACM on Web Conference 2024, pages 103–111, 2024.
介绍了 SAID 框架，该框架利用大型语言模型明确学习基于文本的项目 ID 嵌入的语义对齐。这种方法降低了工业场景中的资源需求。

[157] Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations.arXiv preprint arXiv:2402.17152, 2024.
提出了一种新的推荐系统范式，即生成式推荐器，它将推荐问题重新定义为生成式建模框架内的顺序转导任务。使用生成式推荐器，部署模型的复杂度增加了 285 倍，同时减少了推理计算量。

[30]Breakingthe length barrier: Llm-enhanced ctr prediction in long textual user behaviors. In Proceedings of the 47th InternationalACM SIGIR Conference on Research and Development in Information Retrieval, pages 2311–2315, 2024.
认为，在实际应用中部署大型语言模型的一个关键障碍是它们处理长文本用户行为的效率低下。因此，他们提出了行为聚合层次编码（BAHE），将用户行为的编码与行为之间的交互分开，提高了基于大型语言模型的点击率（CTR）模型的效率。

[98]Llm4sbr: A lightweightand effective framework for integrating large language models in session-based recommendation. arXiv preprintarXiv:2402.13840, 2024.
介绍了一个专门为基于会话的推荐（SBR）定制的可扩展的两阶段大型语言模型增强框架（LLM4SBR）。它研究了将大型语言模型与 SBR 模型整合的潜力，同时兼顾有效性和效率。在涉及短序列数据的场景中，大型语言模型可以利用其语言理解能力直接推断偏好，无需进行微调。这是首次为 SBR 提出大型语言模型增强框架的工作。

加速

在基于LLM的推荐系统领域，加速技术对于优化性能和降低延迟至关重要。**鉴于LLM所需的大量计算资源，提高其部署效率至关重要。**

[142]A decoding acceleration framework for industrial deployable llm-based recommender systems. arXiv preprintarXiv:2408.05676, 2024.
指出了在部署基于大型语言模型的推荐系统时知识生成的低效性，并提出了 DARE。它首次将推测性解码整合到基于大型语言模型的推荐中，从而推动了大型语言模型在推荐系统中的部署。这项工作发现了推荐系统中推测性解码的两个关键特征，并实施了两项增强措施：定制检索池以提高检索效率，以及放宽验证以增加接受的草稿tokens数量。它已在大规模商业环境中的在线广告场景中进行了部署，实现了 3.45 倍的加速，同时保持了相当的下游性能。

冷启动

冷启动是推荐系统中最具挑战性的问题之一。LLM拥有丰富的世界知识，能够更好地理解产品描述中的语义信息以及以文本形式描述的用户偏好信息。因此，**结合LLM有潜力缓解冷启动问题。**

[31] An unified search andrecommendation foundation model for cold-start scenario. In Proceedings of the 32nd ACM International Conferenceon Information and Knowledge Management, page 4595–4601, 2023.
提出了 S&R 多领域基础框架，该框架利用大型语言模型提取领域不变特征，并使用方面门控融合来组合 ID 特征、领域不变文本特征和任务特定的异构稀疏特征，以获得查询和项目的表示。此外，还使用领域自适应多任务模块对来自多个搜索和推荐场景的样本进行联合训练，创建多领域基础模型。他们使用预训练和微调方法将 S&R 多领域基础模型应用于冷启动场景，取得了优于其他最先进的转移学习方法的性能。当代推荐系统主要依赖协同过滤技术，但这种方法忽略了嵌入在项目文本描述中的大量语义信息，导致在冷启动场景中的性能不佳。

[48]Knowledge adaptation from large language model to recommendation for practical industrial application. arXivpreprint arXiv:2405.03988, 2024.
提出了大型语言模型驱动的知识自适应推荐（LEARN）框架，将大型语言模型中包含的开放世界知识有效地整合到推荐系统中。这种方法显著提高了冷启动产品的收入和 AUC 性能。这些改进归因于 LEARN 为购买历史稀疏的产品生成的稳健表示。

[170]Breaking the barrier: Utilizing large language modelsfor industrial recommendation systems through an inferential knowledge graph. arXiv preprint arXiv:2402.13750,2024.
提出，传统的深度点击率（CTR）预测模型通过深度神经网络利用特征交互技术，已在推荐任务中广泛应用。然而，这些模型严重依赖曝光样本和用户反馈，限制了推荐系统在冷启动场景中的性能，并且难以处理新物品的不断出现。他们首次利用大型语言模型的推理能力作为媒介，在向每个用户推荐产品时增强场景偏好，实现了大型语言模型在工业场景中的大规模应用。

动态更新

在工业环境中，基于LLMs的推荐系统的动态更新至关重要，因为它们通过不断适应新的用户行为、内容和趋势，确保推荐保持相关性和准确性。与可能很快过时的静态模型不同，动态更新的模型可以实时或接近实时地响应用户交互和偏好的变化，这在数据快速变化的环境中

至关重要。这种持续的适应性通过提供及时和个性化的推荐来增强用户体验。在高容量、快节奏的工业应用中，动态更新使模型能够从流数据中学习，捕捉最近的用户交互和不断演变的偏好。这种能力不仅提高了推荐的相关性，还使企业能够快速适应用户行为的变化，保持竞争力并优化用户参与度。由于每分钟都有新的内容和产品不断涌入，时间信息对于理解用户偏好以及用户-项目交互随时间的演变至关重要。**在动态推荐系统中整合LLMs在很大程度上仍未被探索，主要是因为将LLMs适应于预测动态变化的数据的复杂性。**

[172] Dynllm:When large language models meet dynamic graph recommendation. arXiv preprint arXiv:2405.07580, 2024.
首次利用连续时间动态图框架将大型语言模型与动态推荐整合在一起。这种整合为动态建模用户偏好和用户-项目交互提供了一种新的视角。他们基于连续时间动态图提出了一种新的大型语言模型增强的动态推荐任务，并提出了 DynLLM 模型，该模型有效地将大型语言模型增强的信息与时间图数据整合在一起。

[157]Actions speak louder than words: Trillion-parameter sequential transducers for generative recommendations.arXiv preprint arXiv:2402.17152, 2024.
提出了 HSTU 架构，专为高基数、非平稳流推荐数据而设计。他们首先将排序和检索任务转换为序列化预测任务，并提出逐点聚合注意力来改进原始的 Transformer 模型。它还包含针对加速推理和减少内存使用的优化解决方案。该解决方案已在拥有数十亿用户的大型互联网平台上部署。

[154]Cosmo: A large-scale e-commerce common sense knowledge generation and serving systemat amazon. In Companion of the 2024 International Conference on Management of Data, pages 148–160, 2024.
展示了 COSMO，这是一个可扩展的系统，旨在从广泛的行为数据中提取以用户为中心的常识知识，并构建行业规模的知识图谱，从而增强各种在线服务。最后，COSMO 已在亚马逊的各种搜索应用中部署。部署围绕高效的特征存储和异步缓存存储展开，确保客户查询和模型响应的流畅处理和成本效益管理。

业务定制需求

在工业应用中，推荐系统需要根据不同业务的独特需求进行定制，每个业务都有其独特的用户群体、内容类型和目标。一刀切的方法是不够的。**例如，电子商务平台需要能够处理多样化的产品、季节性趋势和动态定价的系统，而媒体流媒体服务则优先考虑内容消费模式和观众参与度。能够适应这些特定情境、用户行为和领域知识的可定制模型对于优化性能和实现特定业务目标至关重要，例如提高参与度、转化率或留存率。**基于LLMs的推荐系统的最新进展展示了这种定制的巨大潜力。通过微调这些模型以理解特定用户行为并整合特定领域和外部知识，企业可以创建更具适应性和情境感知的系统。这种灵活性使推荐引擎能够提供更相关和有效的建议，提升用户体验，并更好地与业务目标保持一致。

[84]Modeling user viewing flow using large language models for article recommendation. In Companion Proceedings ofthe ACM on Web Conference 2024, pages 83–92, 2024.
提出了一种名为 SINGLE 的文章推荐任务中的用户视图流建模方法，它包括两部分：恒定视图流建模和瞬时视图流建模。首先，他们使用大型语言模型从之前点击的文章中捕捉恒定的用户偏好。然后，他们通过利用用户文章点击历史与候选文章之间的交互来对用户的瞬时视图流进行建模。

[96]Ad recommendation in a collapsed and entangled world. arXiv preprint arXiv:2403.00793, 2024.
提出了一种工业广告推荐系统，该系统解决了顺序特征、数值特征、预训练嵌入特征和稀疏 ID 特征等问题。此外，他们还提出了有效应对与特征表示相关的两个关键挑战的方法：嵌入维度坍缩和不同任务或场景下的兴趣纠缠。他们探索了几种训练技术，以促进模型优化、减少偏差并增强探索。此外，他们还引入了三个分析工具，允许对特征相关性、维度坍缩和兴趣纠缠进行全面调查。

[158] Notellm: A retrievable largelanguage model for note recommendation. In Proceedings of the 33rd International Conference on World Wide Web,pages 123–132, 2024.
提出了一种新的统一框架 NoteLLM，该框架利用大型语言模型来解决笔记的 I2I 推荐问题。他们使用笔记压缩提示将笔记压缩为一个独特的单词，并通过对比学习方法进一步学习潜在相关笔记的嵌入。推荐系统需要将语义信息与行为数据相结合。当前的主流方法是使用用户和项目 ID 嵌入来增强推荐性能，但这些嵌入往往无法捕捉项目本身的内容相关性，特别是在冷启动问题和基于相似性的推荐场景中。

[3]Beyond labels: Leveraging deep learning and llms for contentmetadata. In Proceedings of the 17th ACM Conference on Recommender Systems, pages 123–145, 2023
讨论了电影推荐系统中内容元数据的重要性，特别是流派标签在理解用户偏好和提供个性化推荐方面的作用。他们指出了使用流派标签相关的挑战，例如流派定义不一致、流派标签的主观性、混合流派的存在以及流派标签无法捕捉视频中流派的强度或程度等。

[88]Trawl: External knowledge-enhanced recommendationwith llm assistance. arXiv preprint arXiv:2403.06642, 2024.
介绍了一种名为 TRAWL（利用大型语言模型增强的外部知识推荐）的推荐系统方法。TRAWL 利用大型语言模型从原始外部数据中提取与推荐相关的知识，并使用对比学习策略进行适配器训练，以增强基于行为的推荐系统。

[153]Heterogeneous knowledge fusion: Anovel approach for personalized recommendation via llm. In Proceedings of the 17th ACM Conference on RecommenderSystems, page 599–601, 2023.
强调了在推荐系统中分析和挖掘用户异构行为的重要性。他们提出了 HKFR，该方法利用大型语言模型从用户行为中提取和整合异构知识，以实现个性化推荐。通过对大型语言模型进行指令调优，并将异构知识与推荐任务相结合，他们显著提高了推荐性能。

[154]Cosmo: A large-scale e-commerce common sense knowledge generation and serving systemat amazon. In Companion of the 2024 International Conference on Management of Data, pages 148–160, 2024.
提出了一种在精心策划的电子商务注释数据上进行微调的语言模型（COSMO-LM），这些数据被组织成指令，以产生与人类偏好一致的高质量常识知识。为了获得大规模和多样化的指令数据，开发了一个自动化的指令生成管道，利用了大量的用户行为。通过扩展产品领域、关系类型和微调任务，该方法实现了可扩展的知识提取。

行业代表

行业代表模型/论文

模型/论文	公司	任务/领域	亮点
LLM-KERec	蚂蚁集团	电子商务推荐	通过LLM构建互补知识图谱
LSVCR	快手	视频推荐	顺序推荐模型和补充的LLM推荐器
RecGPT	快手	顺序推荐	使用个性化提示通过 ChatGPT 对用户行为序列进行建模
SAID	蚂蚁集团	顺序推荐	基于文本利用LLM显式学习语义对齐的项目 ID 嵌入
BAHE	蚂蚁集团	CTR 预测	使用LLM的浅层进行用户行为嵌入，深层进行行为交互
Qiao et al. [98]	华为	基于会话的推荐	在短序列数据中，LLM可以直接利用其语言理解能力推断偏好，而无需微调
DARE	华为	CTR 预测	解决了部署基于LLM的推荐时的推理效率问题，并引入了推测性解码来加速推荐知识的生成
GongGong et al. [31]	蚂蚁集团	多领域推荐	将LLM应用于 S&R 多领域基础模型，以提取领域不变的文本特征
LEARN	快手	顺序推荐	将LLM封装的开放世界知识集成到推荐系统中
Zhao et al. [172]	阿里巴巴	电子商务推荐	根据文本历史购买记录生成用户档案，并通过LLM获得用户嵌入
HSTU	Meta	生成用户动作序列	探索推荐系统的扩展规律；优化模型架构以加速推理
COSMO	亚马逊	语义相关性和基于会话的推荐	是第一个采用LLM构建高质量知识图谱并服务于在线应用的行业规模知识系统
SINGLE	淘宝，阿里巴巴	文章推荐	通过 gpt-3.5-turbo 或 ChatGLM-6B 总结长文章和用户从浏览历史中不变的偏好
Pan et al. [96]	腾讯	广告推荐	通过LLM获得用户或项目嵌入
NoteLLM	小红书	项目到项目的笔记推荐	通过 LLaMA-2 获得文章嵌入并生成标签 / 类别信息
Genre Spectrum	Tubi	电影和电视剧推荐	通过LLM获得内容元数据嵌入
TRAWL	微信，腾讯	文章推荐	使用 Qwen1.5-7B 从文章中提取知识
HKFR	美团	餐饮推荐	使用异构知识融合进行推荐

END