智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

大语言模型及其个性化推荐研究

吴国栋, 秦辉, 胡全兴, 王雪妮, 吴贞畅

引用本文:
吴国栋, 秦辉, 胡全兴, 等. 大语言模型及其个性化推荐研究[J]. 智能系统学报, 2024, 19(6): 1351-1365.
WU Guodong, QIN Hui, HU Quanxing, et al. Research on large language models and personalized recommendation[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(6): 1351-1365.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202309036

DOI: 10.11992/tis.202309036
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240904.1410.002

大语言模型及其个性化推荐研究

吴国栋, 秦辉, 胡全兴, 王雪妮, 吴贞畅
(安徽农业大学 信息与计算机学院, 安徽 合肥 230036)

摘 要: 大语言模型因其强大的自然语言处理能力在人工智能领域产生了巨大影响, 使得大语言模型个性化推荐成为当前推荐系统研究的新兴领域。本文在深入分析已有大语言模型及其个性化推荐相关研究基础上, 探讨大语言模型推荐的过程, 并从直接推荐、基于表示学习推荐、基于生成性学习推荐和提示学习推荐四方面详细分析了大语言模型推荐主要的研究进展。指出现有大语言模型推荐研究中存在的推荐偏差、提示脆弱性、有限上下文、高延迟、公平性和评估等问题, 展望未来大语言模型推荐研究的主要方向, 包括大语言模型推荐的安全性、面向领域的大语言模型推荐、跨模态大语言模型推荐、融合检索任务的大语言模型推荐以及大语言模型推荐的可解释性等。

关键词: 大语言模型; 推荐; 深度学习; 监督微调; 对齐; 提示学习; 生成性; 多模态
中图分类号: TP301 **文献标志码**: A **文章编号**: 1673−4785(2024)06−1351−15

中文引用格式: 吴国栋, 秦辉, 胡全兴, 等. 大语言模型及其个性化推荐研究 [J]. 智能系统学报, 2024, 19(6): 1351–1365.
英文引用格式: WU Guodong, QIN Hui, HU Quanxing, et al. Research on large language models and personalized recommendation[J]. CAAI transactions on intelligent systems, 2024, 19(6): 1351–1365.

Research on large language models and personalized recommendation

WU Guodong, QIN Hui, HU Quanxing, WANG Xueni, WU Zhenchang
(School of Information and Computer, Anhui Agricultural University, Hefei 230036, China)

Abstract: Large language models have revolutionized natural language processing within artificial intelligence, significantly advancing personalized recommendation systems. This paper provides an in-depth analysis of existing research on large language models and their application in personalized recommendations. It explores the process of large language model recommendation and thoroughly analyzes the main research advancements from four perspectives: direct recommendation, representation learning-based recommendation, generation-based recommendation, and prompt learning recommendation. The study identifies several challenges in current research on large language model recommendation, including recommendation bias, vulnerability to prompts, limited contextual understanding, high latency, fairness issues, and evaluation difficulties. It also presents future directions for research on large language model recommendation, including enhancing the security of large language model recommendations, developing domain-oriented large language model recommendations, exploring cross-modal large language model recommendations, integrating retrieval tasks with large language model recommendations, and improving the interpretability of large language model recommendations.
Keywords: large language model; recommendation; deep learning; supervised fine-tuning; alignment; prompt learning; generative; multimodal

大语言模型在个性化推荐领域具有广泛的研究价值和巨大的应用潜力。个性化推荐的目标是根据用户的兴趣和偏好, 为其提供定制化的推荐内容。然而, 传统的个性化推荐方法在处理冷启

收稿日期: 2023−09−21. **网络出版日期**: 2024−09−04.
基金项目: 国家自然科学基金项目 (32371993); 安徽省自然科学基金项目 (2108085MF209); 安徽省科技重大专项 (202103b06020013).
通信作者: 吴国栋. E-mail: gdwu1120@qq.com.

动问题、数据稀疏性和长尾物品推荐方面仍存在一定的挑战。以往的研究工作主要依赖用户行为数据和物品特征信息来进行推荐，但这些方法在捕捉用户意图和语义方面存在一定的限制。随着大语言模型的出现，可以运用深度学习和监督微调等技术，使其具备强大的语义理解和生成能力，为个性化推荐研究带来了新的机遇。

1　大语言模型

大语言模型（large language model, LLM）是一种人工智能模型，旨在理解和生成人类语言。其运用深度学习架构如 Transformer[1] 等技术，在大量文本数据上进行训练，具备强大的语言处理能力。LLM 的特点之一是参数量庞大，通常涵盖数十亿甚至千亿以上的参数，这让大语言模型能够学习和模拟语言数据中的复杂模式，从而在各种自然语言处理（natural language processing, NLP）任务上展现出令人瞩目的表现。训练大语言模型需要大量的文本数据作为语料[2]，这些数据可以涵盖从网站、书籍到文章等各种来源，如表 1 所示。

LLM 训练过程就是通过自动调整模型里的每一个参数来完成这些海量文字的续写，将文本的知识存储在神经网络的参数中。

表 1　主要大语言模型训练数据集

Table 1　Major training datasets for large language models

数据集	介绍
Common Crawl	网络爬虫开放数据库，数据包含原始网页、元数据和文本提取。
网络文本	各种网站、论坛、博客、新闻站点等获取的公开文本内容。
对话数据	人与人之间的对话、问题回答网站上的问答等。
书籍和文学作品	小说、散文、诗歌等各种类型的书籍和文学作品。
学术论文	各个学科领域的学术期刊和会议论文中提取的知识和信息。
维基百科	一个广泛的知识库，其中包含了各种领域的信息和概念。

LLM 训练过程通常分为 4 个主要步骤：预训练、监督微调、奖励建模和强化学习。如图 1 GPT 系列训练过程所示。

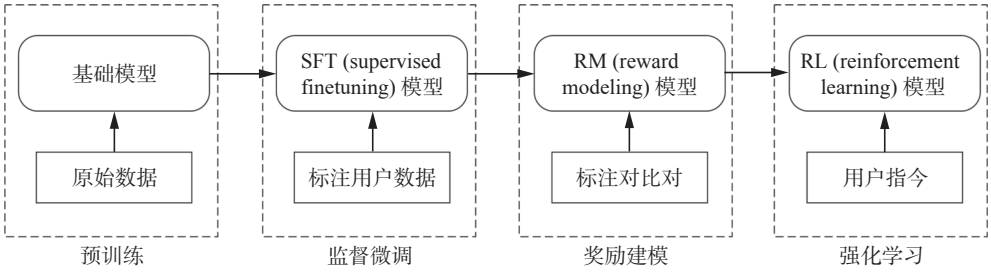


图 1　GPT 系列训练过程

Fig. 1　GPT serious of training processes

在预训练阶段，使用大规模无标注的语料库进行训练，通过学习语料库中的语言模式和结构，模型可以自动学习并表达语言知识。预训练目标是让模型学习到通用的语言知识和表示，为下游任务提供更好的初始化和特征表示。通过自监督学习技术，模型可以预测序列中下一个词或标记，生成自己的标签，并建模之前的词标记。监督微调就是将预训练模型应用到具体任务中。通过在有标注数据上进行微调，模型可以适应特定任务需求，提高模型在该任务上的性能。在监督微调过程中，可以使用不同的损失函数和优化方法，以调整模型的结构和超参数，来优化模型性能。由于训练数据质量不一，大语言模型可能为人类生成有偏见、不符合社会共识的内容。奖励建模目标是让大语言模型输出与人类的价值观和利益相对齐，以使模型生成结果更贴近于人类日常理解习惯，更符合人们所期望的答案。通过收集人类反馈数据，利用强化学习进一步微调大模型，有助于模型微调及理解，并适应特殊任务要求，使模型更加符合人类的偏好。

现有研究已经揭示了大语言模型在上下文学习（in-context learning, ICL）、思维链推理（chain of thought, Toc）和零样本方面卓越的能力[3-5]。上下文学习方面，大型语言模型可以根据对话系统中上下文信息，生成更准确、更自然的文本。此外，模型还可以使用上下文信息进行命名实体识别、关系抽取和情感分析等任务，从而更好地理解和分析自然语言。思维链推理方面，大语言模型可以通过推断和推理，理解和表达自然语言中的逻辑关系与语义关系，如判断逻辑结论真假、推理因果关系及通过思维链推理策略实现循序渐进推理，解决涉及多个推理步骤的复杂任务。零样本方面，大语言模型可以通过学习到的先验知识和关系来进行推理，这些先验知识可以是训练过程

中获得, 并被用来建立一个通用的模型, 使其能够推广到未见过的类别或任务, 为一些特定应用场景带来便利。

2　大语言模型推荐

2.1　大语言模型推荐内涵

大语言模型推荐是指引入深度学习和自然语言处理技术, 利用大语言模型对文本进行理解和生成, 提供个性化、准确和交互性强的推荐结果。大语言推荐模型通过自监督学习在大量数据上训练, 理解用户指令和上下文来增强推荐系统性能。

2.2　大语言模型推荐与传统推荐的主要区别

大语言模型推荐和传统推荐模型都是为了解决推荐系统中用户兴趣多样性、数据稀疏性和冷启动问题。传统推荐模型主要围绕协同过滤、基于内容的过滤和混合推荐等核心技术构建[6-13]。大语言模型推荐相对于传统推荐具有更深层次的语义理解能力、上下文感知性、外部知识覆盖、解释性和交互性及迁移学习能力等优势[14-16]。大语言模型推荐和传统推荐区别主要体现在特征嵌入、模型构建、冷启动、跨域特性、生成特性、外部知识和数据集等上, 如表 2 所示。

表 2　大语言模型推荐和传统推荐主要区别

Table 2　Main differences between large language model recommendation and traditional recommendation

	大语言模型推荐	传统推荐
特征嵌入	将用户和物品的文本描述转化为向量表示	将物品转换为ID, 创建物品嵌入表进行编码
模型构建	预训练语言模型	针对不同任务, 构建不同模型
冷启动	有零/少样本学习能力, 冷启动表现优越	需要足够用户行为数据
跨域特性	捕捉不同领域间的共享知识和语义信息, 能将一个领域的推荐结果迁移到另一个领域	不同领域的间用户兴趣和行为模式存在差异, 领域推荐迁移困难
生成特性	根据用户的输入和上下文信息, 生成符合用户兴趣和需求的推荐结果	无生成特性, 通过多种推荐算法和模型的组合和优化, 为用户提供个性化推荐结果
外部知识	可学习广泛外部知识, 包括常识、语言规则、实体关系等。利用外部知识丰富推荐结果	无外部知识, 通常将物品转换为向量表示, 然后通过比较向量间的相似度来推荐相似物品
数据集	大规模文本数据	用户行为数据

3　大语言模型推荐相关研究

按照对预训练模型学习策略的不同, 现有大语言模型推荐研究可划分为 4 类: 大语言模型直接推荐、基于表示学习的大语言模型推荐、基于生成式学习的大语言模型推荐和基于提示学习的大语言模型推荐。如图 2 所示。推荐系统与大模型交互一般将信息编码为文本注入模型, 并以文本为桥梁建模用户行为。提示学习是大语言模型的重要环节, 基于提示学习的大语言推荐研究分为文本、图结构、多模态 3 种形式, 将在第 4 节详细讨论。

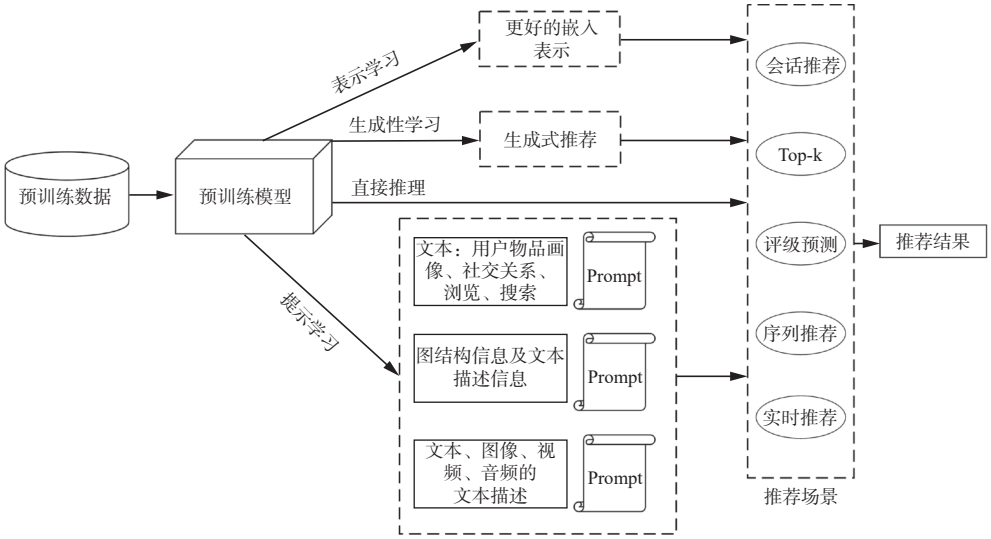


图 2　大语言推荐相关研究范式

Fig. 2　Research paradigms related to large language model recommendation

3.1　语言模型直接推荐

大语言模型直接推荐方式是通过与用户进行对话来获取用户需求和偏好，并基于这些信息生成个性化推荐结果[17-18]，其推荐流程如图 3 所示。

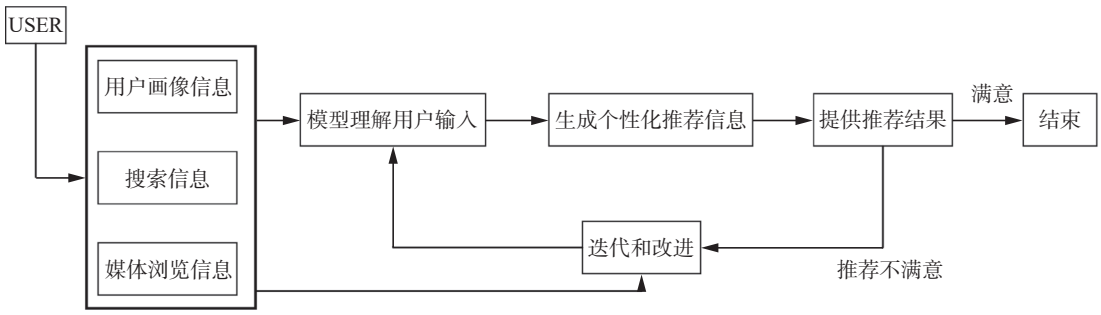


图 3　大语言模型直接推荐流程

Fig. 3　Flowchart of direct recommendation using large language models

1) 用户交互：通过一个对话界面，用户可以进行自然语言与大语言模型的交互。用户可以提供关于他们的需求、偏好和其他相关信息的输入。

2) 理解用户输入：大语言模型需要能够理解用户的输入。它可以使用自然语言处理技术来提取关键信息，如用户的喜好、兴趣领域、时间限制等。这可以通过模型的生成能力和理解能力来实现。

3) 生成推荐信息：基于用户的输入和理解，大语言模型可以根据预训练的知识和上下文生成推荐信息。这可能包括推荐的产品、服务、文章、视频或其他内容。模型可以根据用户的特定需求和偏好进行个性化推荐。

4) 提供推荐结果：大语言模型可以将生成的推荐结果以文本形式返回给用户。这些结果可以包括推荐的物品的摘要、评级、链接等。模型还可以与用户进行进一步的对话，以获取反馈、细化推荐或提供更多选项。

5) 迭代和改进：大语言模型可以继续与用户进行对话，并根据用户的反馈和互动进行迭代和改进。这可以通过模型的生成和理解能力来实现，以提供更准确和个性化的推荐结果。

具体推荐示意如图 4 所示。

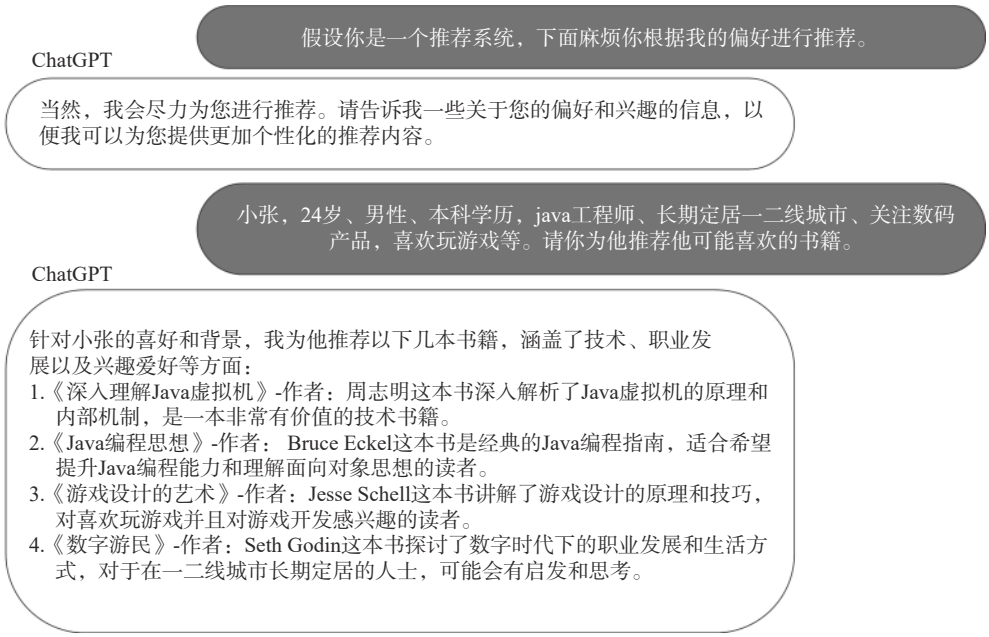


图 4　大语言模型直接推荐示意

Fig. 4　Illustration of direct recommendation using large language models

大语言模型直接能生成针对特定问题或任务的推荐。无论是在聊天应用、社交媒体还是在线购物等场景中，都能够根据用户的即时输入和上下文，快速生成相应的推荐，为用户提供即时的帮助和指导。此外，通过收集用户的反馈信息，大语言模型可以不断学习和改进，从而能够提供更符合用户需求和偏好的推荐。这种个性化、实时性推荐方式将极大地提升用户的使用体验。

3.2　基于表示学习的大语言模型推荐

该推荐方式的核心思想是通过训练语言模型

这种深度学习方法来学习项目、用户和上下文的表示, 然后利用学到的表示来生成个性化的推荐结果。由于推荐系统数据量相对互联网海量文本来说, 规模较小, 数据形式也比较特殊, 所以这种精准调优方式一般基于一个中等规模开源大模型来预训练, 比如 BERT、T5 来进行预训练, 具体可见参考文献 [19-21]。

在特征嵌入方面, 文献 [22] 提出大语言模型推荐利用预训练的语言模型, 如 Transformer 或 BERT 等, 将文本转换为文本表示, 并学习从文本表示到物品表示的转换, 如图 5 所示。这种转换方式能够更好地捕捉物品语义信息, 从而提高推荐准确性。而传统推荐模型通常将物品转换为 ID, 并对物品嵌入进行编码。这种方法无法充分利用文本语义信息, 在某些场景下可能导致推荐效果不佳。

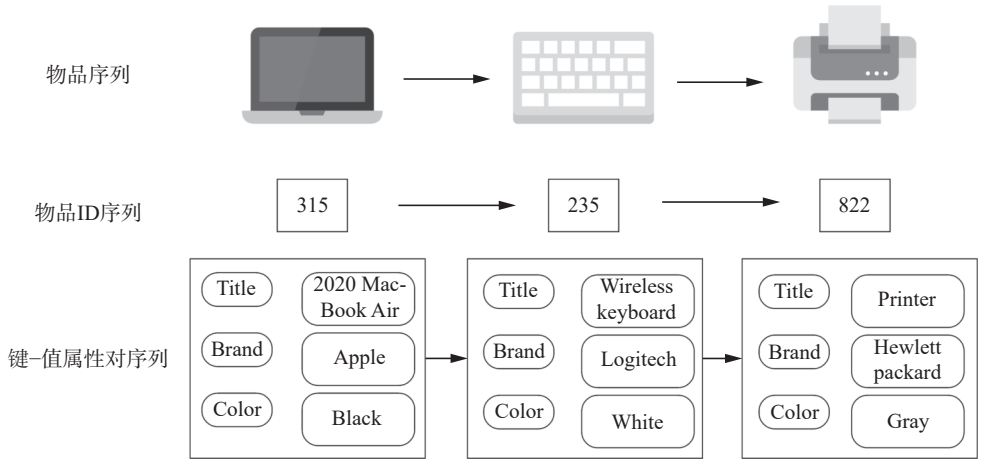
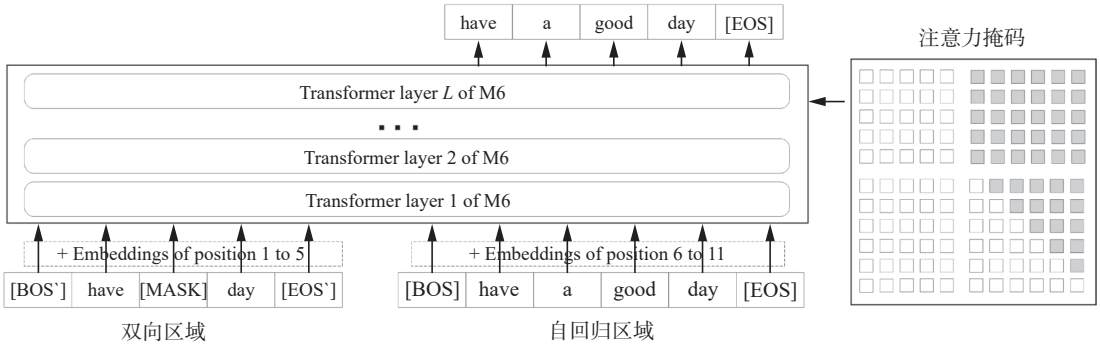


图 5　大语言模型物品嵌入示意

Fig. 5　Illustration of item embedding in large language model

文献 [23] 探讨了基于 M6 的预训练大模型推荐系统。M6-base 包含 30×10^6 参数, 核心结构是 Transformer encoder-decoder, 预训练任务包含完形填空和生成任务, [BOS`] 和 [EOS`] 之间的文本描述了用户的特征, 对应于 M6 输入的第一个区域, 即图 6 中的双向区域。[BOS] 和 [EOS] 之间的文本描述了与候选物品相关的特征, 以及与用户和候选物品都相关的特征, 对应于 M6 输入的第 2 个区域, 即图 6 中的自回归区域。M6-Rec 然后使用 M6 的 Transformer 在 [EOS] 位置的输出向量来总结样本, 将该向量发送到一个线性 softmax 分类器, 并最小化交叉熵损失, 其整体框架如图 6 所示。



注: [BOS] 和 [EOS] 分别表示句子的开头和结尾。

图 6　M6 框架

Fig. 6　M6 framework

基于表示学习的大语言模型推荐能够充分利用大量语言数据和丰富的表示向量, 通过将用户行为数据、物品特征、社交网络数据等融入到表示向量学习中, 模型能够更全面地捕获用户和物品特征和关联性, 提高推荐的准确性和个性化程度, 为大语言模型推荐系统的改进和发展提供了强有力的方法。

3.3　基于生成式学习的大语言模型推荐

传统推荐系统使用的数据局限于一个或少数几个特定应用领域[24-25], 与外部世界的知识隔离, 限制了推荐模型学习的信息。事实上, 超出给定领域的知识可以显著提高推荐系统预测准确性和

泛化能力[26-27]。生成式推荐系统凭借大语言模型中存储的外部知识，直接生成推荐相关内容，不需要计算每个候选分数来进行排序。文献 [28] 指出推荐系统不应仅从封闭系统中狭义数据中学习，而应成为能够主动从外部世界获取知识进行生成式推荐。生成式学习能增强原始数据，扩充数据集，有助于改善模型的冷启动和泛化能力。

文献 [29] 提出了一个新颖而灵活的个性化推荐生成框架。该框架受到搜索引擎启发，首先使用生成式语言模型生成假设的搜索查询，输入包括用户历史中的物品标题和生成提示。然后，通过搜索引擎使用生成的查询来检索推荐物品。这些查询是可理解的，对于解释用户兴趣具有独特的价值，并且使用代表用户兴趣的查询进行推荐，进一步缓解了冷启动问题，其整体框架如图 7 所示。

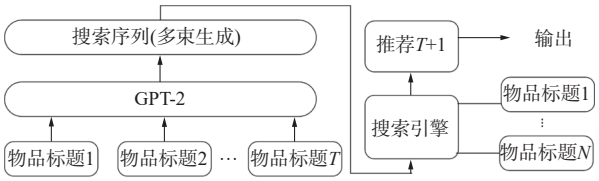


图 7　GPT4Rec 框架

Fig. 7　GPT4Rec framework diagram

文献 [30] 探讨利用大型语言模型中存储的丰富外部知识及其强大文本理解与补全能力，提升和修正简历的描述，有助于弥合用户简历和职位描述之间的信息差。为解决简历生成中的少样本问题，文中提出生成对抗网络进行转移表示学习，从而实现更准确的推荐。文献 [31] 探讨了面向用户的个性化信息需求进行内容生成推荐，提出了一个新的生成式推荐范式 GeneRec。文中首先预处理用户指令和传统反馈，作为生成的依赖。基于 AI editor 和 AI creator 来实例化 AI generator，使得 GeneRec 可以基于用户需求重新定制已有物品和创建新的物品，其模型如图 8 所示。

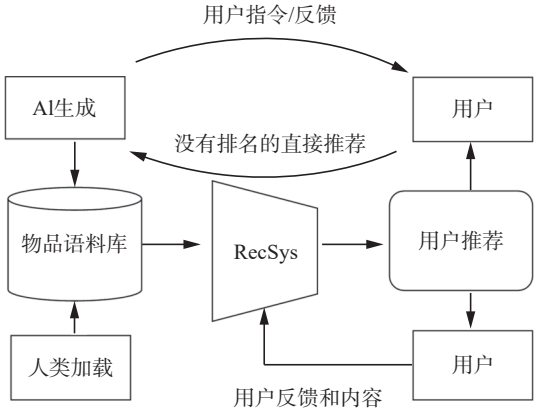


图 8　GeneRec 范例示意

Fig. 8　Illustrative example diagram of GeneRec

基于生成式学习的大语言模型推荐可以生成个性化推荐内容，满足用户需求和兴趣。这种个性化能力使得推荐结果更加精准和符合用户的偏好。生成式推荐不仅可以推荐用户已知的物品，还能推荐出新颖的物品，使用户获得更多选择。此外，生成式推荐还能解决冷启动问题，即使在没有大量用户行为数据或物品信息的情况下，也能生成个性化推荐结果。因此，生成式推荐在推荐方面具有创新性和个性化等特点，能够为推荐系统带来更好的用户体验。

4　基于提示学习的大语言模型推荐

大语言模型虽然具有很强的语言建模能力，但在推荐系统中需要进行提示信息和指令调优才能实现更好的推荐效果。这是因为推荐系统需要考虑用户的历史行为和兴趣，以及物品的属性和内容等多方面因素。通过利用指令调优和提示，没有改变大模型的参数，但激活了大模型神经网络的某个功能区域，以帮助大语言模型更好地理解和捕捉这些因素，从而实现更准确的推荐。

4.1　基于文本的 Prompt

这种方法主要基于用户输入的文本提示，用文本高效表示用户行为。通过大语言模型进行语义分析和关联度计算，从而推荐相关内容。

P5 开创性的工作[32] 说明了将提示制定为自然语言任务的可行性，并对广泛使用的开源 T5 模型进行了细化，以创建一个能够处理各种任务的统一框架。这种创新方法突出了大语言模型在推荐环境中处理多任务学习的多功能性。然而，大语言模型理解和生成基于文本推荐的潜力尚未得到充分探索。文献 [33] 提出了一种基于文本生成推荐的新方法，目标是解决以前工作的一些局限性，并突破推荐系统领域可能范围。该模型利用大语言模型的理解能力来解释上下文、学习用户偏好，并利用大型语言模型参数中大量知识来完成推荐任务。首先制定了专门提示，以增强大语言模型理解推荐任务的能力；随后，使用文本数据表示的用户−物品交互提示，微调基于 LLaMA 骨干的大语言模型，以捕获用户偏好和物品特征。文献 [34] 对这一新范式进行了首次尝试，开发了一个新闻推荐的提示学习框架 (Prompt4NR)，该框架将预测用户是否会点击候选新闻的任务转换为完形掩码预测任务，并设计了一系列提示模板，包括离散模板、连续模板和混合模板，并构建了相应的答案空间，以检验所提出的 Prompt4NR 框架，其具体步骤如下：

1)在 Prompt4NR 数据格式转换模块中, 点击历史记录和候选新闻被转换为自然语言句子, 分别表示为<USER>和<CANDIDATE>。其中, <USER>是用户兴趣领域的摘要, 通过将点击历史记录中的新闻标题连接在一起得到; <CANDIDATE>是候选新闻的标题。

2)提示模板模块是一组基于语义相关性、用户情感、用户行为和推荐效用 4 方面设计的模板。作者设计离散、连续和混合 3 种类型, 并为每个类型模板构建了一个答案空间, 如表 3 所示, 其中 $[P_1]\cdots[P_{n_1}]$、$[Q_1]\cdots[Q_{n_2}]$ 和 $[M_1]\cdots[M_{n_3}]$ 表示虚拟可学习的标记。

表 3　Prompt 模板设计, 包括离散、连续和混合模板

Table 3　Prompt template design, including discrete, continuous, and mixed templates

类型	视角	模板T(<USER>,<CANDIDATE>,[MASK])	回复词
离散样本	语义相关性	<CANDIDATE>对于<USER>来说是[MASK]的。	相关/不相关
	用户情感	根据用户的兴趣领域<USER>, 他对<CANDIDATE>的感受是[MASK]。	感兴趣/无聊
	用户行为	用户: <USER>, 新闻: <CANDIDATE>。用户是否点击这条新闻? [MASK]。	是/否
	推荐效用	根据<USER>, 向用户推荐<CANDIDATE>是一个[MASK]的选择。	好/坏
连续样本	语义相关性	$[Q_1]\cdots[Q_{n_2}]$<CANDIDATE> $[M_1]\cdots[M_{n_3}]$[MASK] $[P_1]\cdots[P_{n_1}]$ <USER>	相关/不相关
	用户情感	$[M_1]\cdots[M_{n_3}]$ [MASK] $[Q_1]\cdots[Q_{n_2}]$ <CANDIDATE> $[P_1]\cdots[P_{n_1}]$ <USER>	感兴趣/无聊
	用户行为	$[P_1]\cdots[P_{n_1}]$ <USER> [SEP] $[Q_1]\cdots[Q_{n_2}]$ <CANDIDATE> [SEP] $[M_1]\cdots[M_{n_3}]$ [MASK]	是/否
	推荐效用	$[Q_1]\cdots[Q_{n_2}]$ <CANDIDATE> $[M_1]\cdots[M_{n_3}]$ [MASK] $[P_1]\cdots[P_{n_1}]$ <USER>	好/坏
混合样本	语义相关性	$[P_1]\cdots[P_{n_1}]$ <USER> [SEP] $[Q_1]\cdots[Q_{n_2}]$ <CANDIDATE> [SEP] 这条新闻对于用户的兴趣领域来说是[MASK]的。	相关/不相关
	用户情感	$[P_1]\cdots[P_{n_1}]$ <USER> [SEP] $[Q_1]\cdots[Q_{n_2}]$ <CANDIDATE> [SEP] 用户对这条新闻的感受是[MASK]。	感兴趣/无聊
	用户行为	$[P_1]\cdots[P_{n_1}]$ <USER> [SEP] $[Q_1]\cdots[Q_{n_2}]$ <CANDIDATE> [SEP] 用户是否点击了这条新闻? [MASK]。	是/否
	推荐效用	$[P_1]\cdots[P_{n_1}]$ <USER> [SEP] $[Q_1]\cdots[Q_{n_2}]$ <CANDIDATE> [SEP] 向用户推荐这条新闻是一个[MASK]的选择。	好/坏

3)在答案预测的标记转自然语言模块中, 作者使用包含两个具有相反含义的词语的二元答案空间, 将推荐标签映射到回答 [MASK] 预测的特定单词, Prompt4NR 整体框架如图 9 所示。

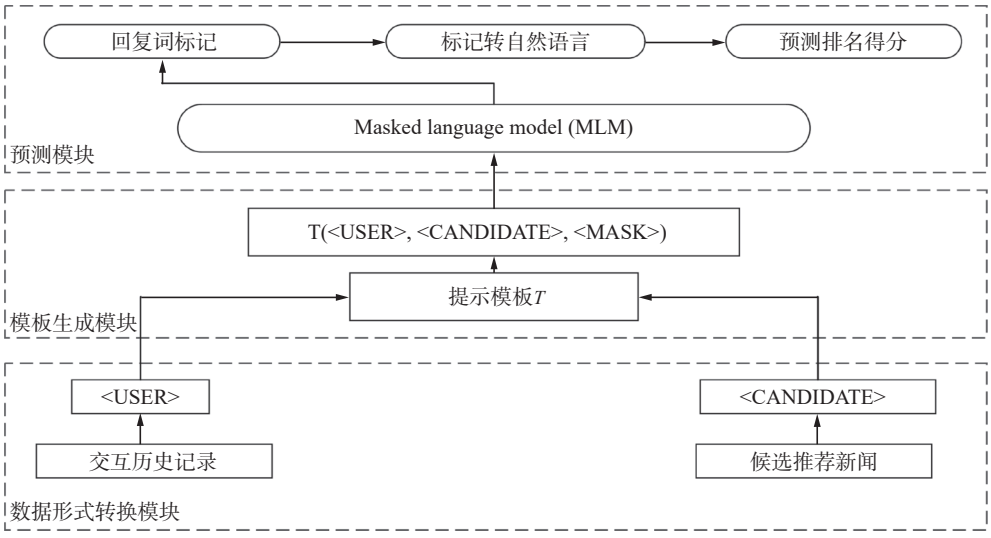


图 9　Prompt4NR 框架

Fig. 9　Prompt4NR framework

4.2　基于图结构的 Prompt

基于图的大语言模型推荐是推荐系统一种新兴的研究方向，其利用了图结构数据和大语言模型的强大能力来实现更有效和更个性化推荐。该方法基本思想是将用户、物品和上下文信息表示为图中的节点和边，然后使用大语言模型对这些节点和边进行嵌入学习，从而捕捉它们之间的复杂关系和语义信息。基于这些嵌入，可以使用合适的推荐模型来预测用户对物品的兴趣和行为，并生成个性化推荐结果。大语言模型主要优势在于自然语言处理能力。相比之下，图数据通常需要使用特定的技术和算法进行处理分析。因此，大语言模型在处理图数据方面的能力可能有限。然而，最近的一些研究表明，大语言模型确实可以通过结合知识图谱或者用户行为图等图结构数据来更好地理解和处理图数据。如可以将自然语言文本和图数据一起输入到模型中，学习到自然语言描述和图结构之间的映射关系。

对于大语言模型推荐来说，基于图数据的任务可以根据其目标分为两类。第 1 类是结构理解任务，如识别图中重要节点、计算中心度指标[35-37]等。第 2 类是语义理解任务，如知识图谱问答[38-39]、节点分类[40]和图分类[41]等，这些任务具有不同的需求和挑战。

文献 [42] 探讨了大语言模型是否能解决自然语言中的图问题。文献 [43] 提出了一个在大型图语料库上进行图感知语言模型预训练框架，它结合了大语言模型和图神经网络，并在下游应用程序上采用各种微调方法。鉴于图结构数据在社交网络分析、药物发现、推荐系统和时空预测等众多应用中的普遍存在和不可或缺的作用，文献 [44] 开始探索大语言模型是否能够理解图形结构化数据，并进行了实证评估和基准测试。该文献首先提出了一个集成大型语言模型和图结构数据的新框架，以增强它们在广泛的图挖掘任务中的协同能力，从而弥合大型语言模型之间现有差距，如图 10 所示。基于上述框架，提出 10 种常见场景来评估语言模型在处理图相关任务方面的能力。

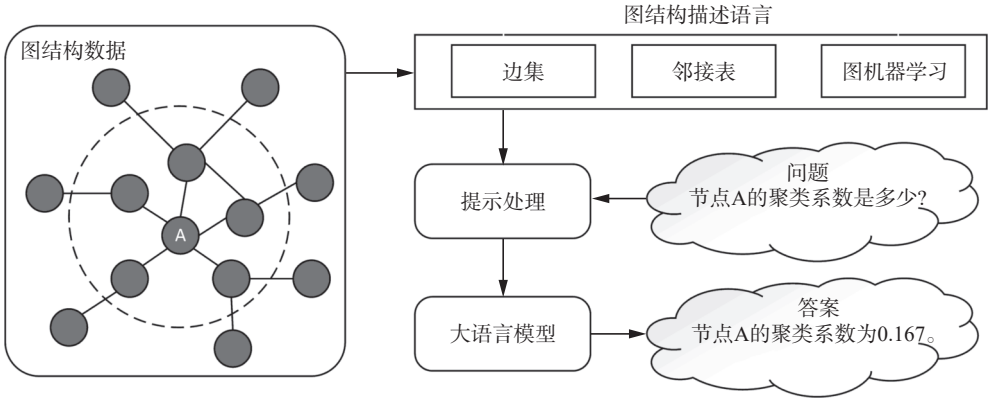


图 10　大语言模型框架图形理解示意

Fig. 10　Illustrative diagram of the framework for visual understanding in large language models

文献 [45] 探索大语言模型在行为图理解中的潜力，提出了利用元路径提示构造器将非文本行为图信息编码为自然语言提示的方法，并利用这个元路径提示来增强在线招聘推荐的效果。如图 11 所示。职位推荐具体步骤如下：

1）构建异构图。求职者和职位之间的交互涉及不同类型的行为，形成了一个行为图。这个行为图是一个典型的异构图，其中不同的节点类型包括候选人、职位，不同的边类型包括消息、面试、匹配等。

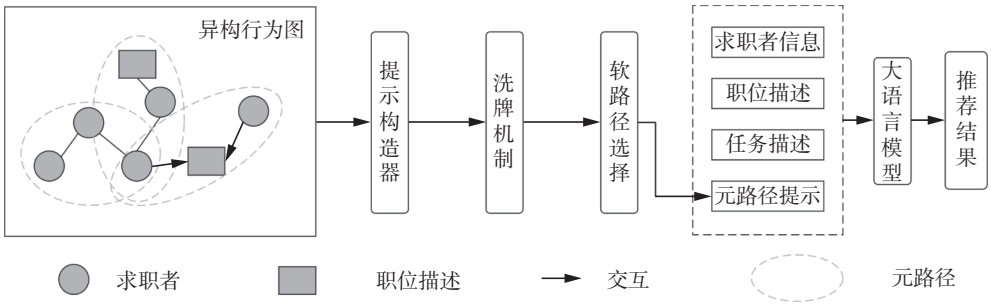


图 11　职位推荐框架

Fig. 11　Framework diagram for job recommendation

2)预定义提示模板。使用简历或职位描述信息填充模板。由于行为图中每种类型的边具有独特和定义明确的语义, 因此自然而然地考虑将图数据格式中的元路径转换为大语言模型可接受的自然语言描述。

3)相似度计算。为了避免相似的元路径导致冗余, 定义了一个简单的相似度度量:

$$S_{i,j}=\frac{\left|P_i \cap P_j\right|}{\left|P_i \cup P_j\right|},\ P_i,\ P_j \in \Phi_p \tag{1}$$

式中: Φ_p 表示候选项的采样元路径集合; P_i 和 P_j 表示该集合中的两个元路径; $|P_i \cap P_j|$ 表示同时存在于两个路径中的标记数, 而 $P_i \cup P_j$ 则表示它们的并集。

4)处理元路径。由于不同的路径将为模型决策带来不同的权重, 路径提示的顺序偏差会带来不稳定的答案。为了解决这个问题, 设计了路径随机化、自适应路径选择和它们的混合路径增强机制, 以减轻不同路径提示带来的负面影响。

5)调整指令。通过最小化由真实标签和相应的大语言模型输出计算得出的自回归损失来进行指令微调。在实验中, 屏蔽了提示的损失位置, 采用了特定的提示格式、任务特定的指令和真实标签。最终的学习目标可以计算:

$$\mathcal{L}_f=\max_{\theta_L}\sum_{(x,y)\in\mathcal{T}}\sum_{t=1}^{|\mathcal{Y}|}\log(P_{\theta+\theta_L}(y_t|e_x,y<t)) \tag{2}$$

其中 θ_L 是 LoRA 的参数。

6)进行推荐。由于经过多次对齐步骤后, 训练好的模型已经学习了定义的真实标签的输出格式。模型输出中捕获标签生成的 softmax 概率(用于表示标签的令牌, 例如本文中的“Yes./No.”或“[A]/[B]”), 并在模型输出中对应于真实标签位置进行计算, 得出最终预测概率。

总之, 虽然大语言模型可能不是最适合处理图结构数据的工具, 但通过结合自然语言文本, 它可以在一定程度上理解和处理图数据。未来, 随着深度学习技术的不断发展, 大语言模型在处理图结构数据方面的能力可能会得到进一步提升。

4.3　基于多模态的 Prompt

基于多模态的大语言模型推荐是指可以处理、分析和交互多种类型数据的推荐系统, 不仅可以能够处理单一数据类型任务, 而且可以在不同数据类型间建立联系和融合, 从而实现一个综合、全面的理解。

一种常见的方法是将多模态数据编码为文本表示形式, 然后将其输入到大语言模型中进行训练和推荐。例如, 可以使用图像或视频的标注文本作为输入, 然后使用预训练的大语言模型对其进行编码, 最后使用推荐算法将编码向量与用户或项目进行匹配。这种方法的优点是可以使用现有大量文本数据进行预训练, 并且可以使用通用文本推荐算法进行实现。

另一种方法是使用深度学习模型直接对多模态数据进行建模。例如, 可以使用深度神经网络对图像或视频进行特征提取, 然后将其与文本数据组合在一起, 构建一个多模态表示形式; 最后, 可以使用大语言模型对这种多模态表示进行训练和推荐。这种方法优点是可以更好地捕捉多模态数据中的复杂关系和交互, 并且可以通过联合训练来提高模型性能。

MIniGPT4[46] 使用图像描述数据将视觉编码器和语言模型进行对齐, 从而实现赋予语言模型视觉能力的目标。BLIP2[47] 是较早提出“大语言模型 + 视觉编码器”多模态模型构想的研究, 其整体结构如图 12 所示。该研究主要提出了跨视觉语言模态的 Q-former 连接结构, 如图 13 所示, 其设计包括图像−文本匹配、基于图像的文本生成和图像−文本对比学习等 3 个对齐语言和视觉特征部分。BLIP2 采用了 ViT-L/g 作为图像编码器, 使用的大语言模型则为 OPT 和 Flan T5 语言模型, 这些模型在语言生成方面的能力相对较弱。BLIP2 的预训练分为两个阶段: 第 1 阶段是利用 Q-former 与一个冻结参数的图像编码器进行训练, 以学习视觉和语言的表示; 第 2 阶段是利用 Q-former 与冻结的大语言模型进行训练, 以学习视觉到文本的生成能力。然而, BLIP2 模型的一个缺陷是缺乏上下文学习能力, 因为其训练数据是单个图像−文本对, 缺乏多轮对话的相关性。

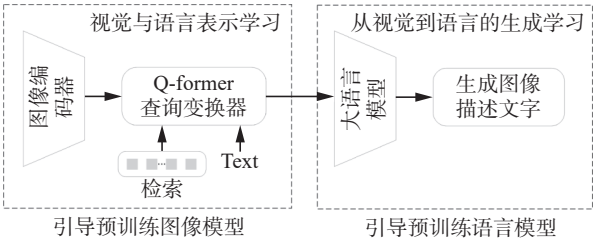


图 12　多模态模型推荐整体

Fig. 12　Overall diagram of multimodal model recommendation

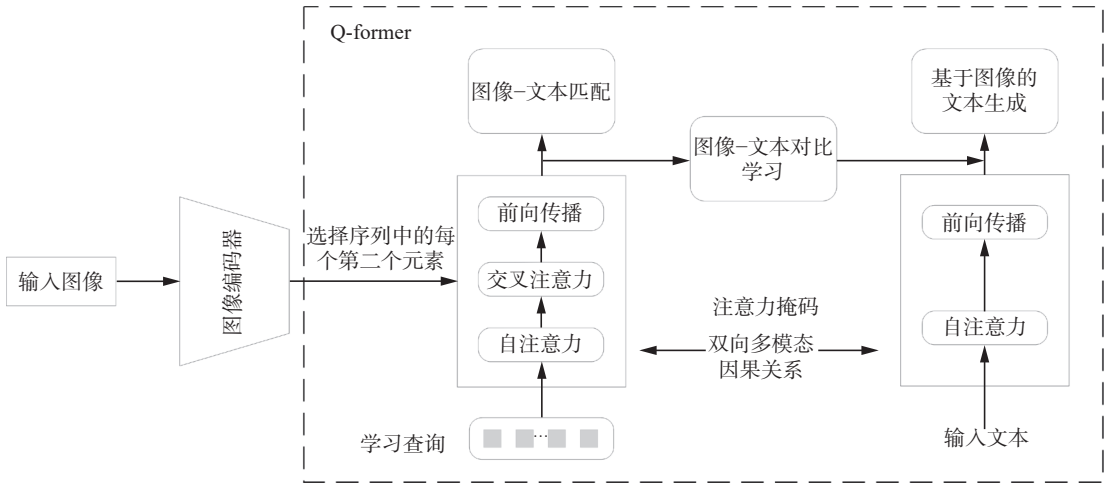


图 13　Q-former 结构

Fig. 13　Q-former architecture

对于多模态大模型的构建，目前主流方法是设计两阶段训练方法。其中第 1 阶段均采用图文建立视觉模型和语言模型之间的联系，这个过程对于多模态大模型是必要的，因为视觉模型和生成式语言模型分别在各自模态预训练，存在特征不匹配问题。第 2 阶段训练模型，则是可以根据数据量和需求进行调整选择。当前融合大语言模型和视觉模型的方式还相对简单，但是已经展现出了优秀的推荐效果，未来多模态通用模型可能成为人工智能的下一个发展目标。

基于提示学习的大语言模型在推荐系统中具有显著优势。它能够理解用户提示信息和推荐任务上下文，从而更准确地满足用户需求和偏好。模型的可解释性使得用户更加理解和信任推荐结果。此外，通过在不同任务和领域中共享提示信息，大语言模型展现出良好的迁移学习和泛化能力[48]，能够更好地利用已有知识和经验，提高推荐效果和效率。

4.4　模型评估和比较分析

为评估不同模型在相同数据集上的性能，进行了一系列相关实验。实验旨在比较各个模型在特定任务上的表现，并提供深入的洞察。通过这些比较，可以了解不同模型的优势和局限性，以便更好地选择适合特定需求的 LLM 推荐方法，并为进一步的研究和应用提供指导。

本实验在 MovieLens 100K 数据集上进行。为了保证实验的准确性，实验对数据集进行了基本的预处理。删除了交互次数少于 10 个的用户和物品。然后，根据 8∶1∶1 将数据集分为训练集、验证集和测试集。为了评估模型，实验采用了两个在之前的工作中广泛使用的评价指标：Recall@10/20 和 NDCG@10/20 来比较不同模型之间的推荐性能，其分别衡量系统在前 10/20 个推荐项中找回相关项的能力和推荐系统排序质量。模型使用 Adam 优化器来优化模型参数。初始学习率和批量大小分别设置为 0.01 和 64。其次，从{0.1, 0.05, 0.01, 0.005, 0.001}的集合中选择具有权值衰减的正则化策略来缓解训练阶段的过拟合问题，如表 4 所示。

表 4　基于 Recall@10/20、NDCG@10/20 的性能比较

Table 4　Performance comparison based on Recall@10/20, NDCG@10/20

方法	Baseline	R@10	N@10	R@20	N@20
传统协同过滤	MF-BPR	0.1890	0.0815	0.2564	0.0985
	NGCF	0.2084	0.0886	0.2926	0.1100
	LightGCN	0.1994	0.0837	0.2660	0.1005
表示学习方法	Recformer	0.1948	0.1027	0.3114	0.1252
	M6Rec	0.2482	0.1113	0.3354	0.1310
生成式方法	Gpt4Rec	0.1152	0.0672	0.1543	0.0788
	GenRec	0.1311	0.0837	0.1614	0.0972

实验结果表明，M6Rec 性能在 MovieLens 100 K 数据集上有显著的提升。主要原因是 M6Rec 考虑了用户物品文本嵌入信息。生成式推荐方法推荐效率下降的主要原因是生成式推荐方法通常会注重推荐的多样性和创造性，以提供用户新的推荐体验。然而，为了实现多样性和创造性，推荐系统需要更多的计算资源和时间来生成和评估各种可能的推荐结果。由于提示学习用的数据集通常是具有文本信息描述的数据集，此外不同模型针对不同的数据集设计不同的推荐提示模板，因此难以在同一个数据集上对提示学习进行评估。

5 大语言模型推荐研究的主要问题

5.1 推荐偏差问题

由于大语言模型推荐生成具有创造性, 大语言模型推荐可能会推荐候选集合中不存在的商品。文献 [49] 研究试图通过将候选商品集合输入到提示中来解决这个问题, 但这种方法也同样面临着位置偏差和上下文长度有限等问题。语言模型的位置偏差问题和流行度偏差问题可能会导致推荐结果不准确或不公平。其中, 位置偏差问题可以通过采用随机抽样法和强调最近的交互项来缓解; 流行度偏差问题比较困难, 因为它与预训练语料库的组成密切相关, 热门的物品在大模型的训练语料库中更多, 导致大模型在应用于推荐时, 也会倾向于推荐热门的物品。

5.2 提示 (prompt) 的脆弱性问题

大语言模型对输入的提示极为敏感, 并且可能不会完全遵循指令提示。文献 [50] 中, 研究人员在输入的提示中添加额外的句子, 比如“在没有解释的情况下, 给出一个数字作为评分”“不要给出推理”等, 以避免大语言模型输出不符合预期的电影评分预测。由于大型语言输入语法、语义的细微变化会导致剧烈的输出变化, 因此了解大语言模型如何对其输入的微小变化做出反应对于确保公平性和实用性至关重要。文献 [51] 引入了一种称为对比输入解码 (contrastive input decoding, CID) 的新方法, 通过生成反映 2 个略有不同输入的独特特征的文本来阐明这些细微差异, 从而使大语言模型的响应更易于理解和管理。CID 作为一种解码算法, 提供了一种有用的工具, 但仍需要更多的研究来解决这些重要问题。

5.3 有限上下文问题

大语言模型中有限的上下文长度使得很难将大量的推荐物品集输入到大模型中, 导致大模型生成不包含在物品集中的物品。其次, 用户行为序列或历史会话可能无法完整输入到模型中, 这使得大模型无法获取用户全面兴趣偏好, 导致推荐精度下降。一些研究人员正在研究如何使用多层次上下文信息来生成更准确和连贯的文本。另外一些研究人员正在探索如何结合知识图谱和常识知识来扩展大语言模型的上下文范围。但这些方法目前没有在工业界得到有效性验证。为了解决这些问题, 需要进一步研究如何扩展上下文数量限制, 以提升大模型的上下文学习和思维链能力[52], 从而改善推荐系统的效果。

5.4 高延迟问题

大语言模型并行度低, 内存需求大, 部署困难。在推荐系统中的应用可能会面临推荐延迟过高的问题。这是因为大语言模型需要处理大量的输入数据, 并进行复杂的计算和推理, 这些过程可能需要较长时间才能完成。在实时推荐系统中, 这种推荐延迟过高可能会导致用户体验下降, 从而影响系统的性能和可用性。

5.5 公平性问题

研究人员发现, 预训练语料库中的偏见可能会误导大语言模型推荐生成有害或冒犯性内容, 例如, 歧视处于劣势群体中的人们。虽然有一些策略[53] 可以减少大语言模型的有害性, 但现有的研究已经从用户方面[54] 和物品方面[3] 发现了大语言模型带来的推荐不公平问题。文献 [55] 提出了一个新的公平性基准“Fairness of Recommendation via LLM (FaiRLLM)”, 具体来说, FaiRLLM 通过比较大语言模型在“中性指令” (没有包含用户的敏感属性) 和“敏感指令” (包含敏感属性) 下的推荐结果, 来评估大语言模型的公平性。结果表明, 大语言模型可能产生不公平的推荐, 而且大语言模型推荐的公平性随着不同的敏感属性而变化。

5.6 性能评估问题

在使用 MovieLens、Amazon Books 等数据集测试大语言模型推荐和零/少样本学习能力时可能存在的问题。因为这些数据集相对于实际工业数据集来说规模比较小, 可能不能充分反映大语言模型推荐能力, 而且这些数据集中的物品可能与大语言模型的预训练数据有关, 这可能会引入评估大语言模型零/少样本学习能力的偏差。传统推荐中 NDCG、MSE、Recall 等用于评估推荐结果与实际用户交互物品之间的差异。针对大语言模型推荐需要消耗大量的算力, 应进一步考虑实时性和计算资源使用情况, 如内存使用、CPU 占用率等。此外, 大语言模型推荐还需考虑可信度问题、领域特定限制和道德等方面, 以确保推荐系统的可靠性和合规性[56]。

6 大语言模型推荐未来主要研究方向

6.1 垂直领域专属大语言模型推荐

目前来看, 公有大模型有很多的不足, 第一是其是个“通才”, 但是缺乏行业深度[57]。ChatGPT 什么都知道, 但是对于行业领域的信息只能泛泛而论, 知识深度不够。将大语言模型的“通才”能力转变为推荐领域的“专才”能力, 需要在数据

集、模型架构和算法等方面进行优化和调整。这将有助于解决大语言模型在推荐领域的偏差和可解释性问题。

6.2　融合多模态信息的大语言模型推荐

推荐系统涉及到的数据很多都是多模态的，比如物品有描述文本，有图片，甚至有介绍的视频等，这些异构的信息对于推荐系统的效果相当重要。当前的大模型以处理文本数据为主，还无法很好地处理多模态数据。因此，研究多种信息进行推荐系统建模的深度学习推荐算法，可以提供更全面和多样化的特征表示，从而增强推荐系统的健壮性和准确性。此外，利用多模态大模型来进行推荐，可以进一步提升模型对于上下文和用户需求的理解，从而更好地应对大语言模型推荐中可能存在的提示脆弱性问题。

6.3　融合检索任务的大语言模型推荐

OpenAI 在 2023 年 3 月宣布在 ChatGPT 中推出 Plugins 插件系统。插件专门为大语言模型设计，通过与开发者定义的 API(application programming interface) 互动，将大模型连接到第三方应用程序，使其能够访问互联网最新信息。实时检索感兴趣的信息在新闻推荐领域扮演着重要的角色。新闻推荐领域旨在利用先进的自然语言处理技术和实时数据源，为户提供最新、个性化的新闻推荐。通过整合实时数据源和新闻聚合网站的 API，大语言模型可以自动地检索和获取最新的新闻。大语言模型推荐通过查询外部知识库，获取与推荐相关的上下文信息，很大程度解决了大语言模型推荐中有限上下文问题，并准确地推荐符合用户兴趣的新闻。

6.4　大语言模型推荐的安全性

大模型推荐系统处理大量用户数据，包括个人偏好、历史行为等敏感信息。需要学术界致力于开发技术和算法，以保护用户隐私并降低个人信息的泄露风险。未来的大模型推荐系统将更多地集成安全性技术。这可能包括差分隐私技术、同态加密、安全多方计算等，以确保在数据处理和分析过程中保持用户数据的隐私安全。同时，技术创新和法律法规的进一步发展也将对数据安全性提出更高要求。

6.5　大语言模型推荐的可解释性

ChatGPT 是一个基于大数据和统计学的语言模型，通过学习海量文本，预测一个概率最高的词。GPT 可以回答大量它没有学习过的知识，比如简单的知识推理。甚至还学会了在对话中临时学习的能力。2023 年 5 月，OpenAI 的新研究提出，可以采用 GPT-4 来进行 GPT-2 的神经元解释[58]。当将文本输入 GPT-2 时，模型中的特定神经元会被激活。随后，OpenAI 引导 GPT-4 观察这一过程，并尝试推断这些神经元的功能。通过进一步观察更多的文本和神经元，GPT-4 能够逐渐推测出 GPT-2 中每个神经元的功能。这一方法有助于我们深入理解 GPT-2 的内部运行机制。对于理解、描述和解释大语言模型推荐能力的正式理论和原理，仍然缺乏更多的了解。这些基本问题值得深入研究和探讨，对于下一代大语言模型推荐占用大量计算资源和公平性问题具有至关重要的意义。

7　结束语

本文聚焦于大语言模型推荐领域，从预训练模型调优策略的视角，探讨了已有大语言模型及其推荐相关研究。通过利用大语言模型对文本、图结构、多模态语义理解和上下文关联能力，发现大语言模型推荐在提供更准确、多样化且用户满意度更高的推荐结果方面具有巨大的优势，有效缓解了传统推荐方法中数据稀疏性和冷启动等问题[59]。最后，对已有大语言模型推荐研究的优点与不足进行了分析，指出了现有大语言模型推荐研究存在的问题及未来主要研究方向，为进一步探讨大语言模型推荐相关研究提供了一定的借鉴意义。

参考文献：

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017−06−12)[2023−08−25]. http://arxiv.org/abs/1706.03762.

[2] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[EB/OL]. (2020−05−28)[2023−08−25]. http://arxiv.org/abs/2005.14165.

[3] HOU Yupeng, ZHANG Junjie, LIN Zihan, et al. Large language models are zero-shot rankers for recommender systems[EB/OL]. (2023−05−15)[2023−08−25]. http://arxiv.org/abs/2305.08845.

[4] WANG Lei, LIM E P. Zero-shot next-item recommendation using large pretrained language models[EB/OL]. (2023−04−06)[2023−08−25]. http://arxiv.org/abs/2304.03153.

[5] KOJIMA T, GU S S, REID M, et al. Large language models are zero-shot reasoners[EB/OL]. (2022−05−24)[2023−08−25]. http://arxiv.org/abs/2205.11916.

[6] HE Xiangnan, LIAO Lizi, ZHANG Hanwang, et al. Neur-

al collaborative filtering[C]//Proceedings of the 26th International Conference on World Wide Web. [S.l.: s.n.], 2017: 173−182.

[7] 黄璐, 林川杰, 何军, 等. 融合主题模型和协同过滤的多样化移动应用推荐[J]. 软件学报, 2017, 28(3): 708−720.
HUANG Lu, LIN Chuanjie, HE Jun, et al. Diversified mobile app recommendation combining topic model and collaborative filtering[J]. Journal of software, 2017, 28(3): 708−720.

[8] SON J, KIM S B. Content-based filtering for recommendation systems using multiattribute networks[J]. Expert systems with applications, 2017, 89: 404−412.

[9] 刘建勋, 石敏, 周栋, 等. 基于主题模型的 Mashup 标签推荐方法[J]. 计算机学报, 2017, 40(2): 520−534.
LIU Jianxun, SHI Min, ZHOU Dong, et al. Topic model based tag recommendation method for mashups[J]. Chinese journal of computers, 2017, 40(2): 520−534.

[10] BASILICO J, HOFMANN T. Unifying collaborative and content-based filtering[C]//Twenty-first international conference on Machine learning. Banff: ACM, 2004: 65−72.

[11] 曹俊豪, 李泽河, 江龙, 等. 一种融合协同过滤和用户属性过滤的混合推荐算法[J]. 电子设计工程, 2018, 26(9): 60−63.
CAO Junhao, LI Zehe, JIANG Long, et al. A hybrid recommendation algorithm based on collaborative filtering and user attribute filtering[J]. Electronic design engineering, 2018, 26(9): 60−63.

[12] 孙冬婷, 何涛, 张福海. 推荐系统中的冷启动问题研究综述[J]. 计算机与现代化, 2012(5): 59−63.
SUN Dongting, HE Tao, ZHANG Fuhai. Survey of cold-start problem in collaborative filtering recommender system[J]. Computer and modernization, 2012(5): 59−63.

[13] DA'U A, SALIM N. Recommendation system based on deep learning methods: a systematic review and new directions[J]. Artificial intelligence review, 2020, 53(4): 2709−2748.

[14] ZHANG Jiawei. Graph-ToolFormer: to empower LLMs with graph reasoning ability via prompt augmented by ChatGPT[EB/OL]. (2023−04−10)[2023−08−25]. http://arxiv.org/abs/2304.11116.

[15] PAPARRIZOS I, CAMBAZOGLU B B, GIONIS A. Machine learned job recommendation[C]//Proceedings of the fifth ACM conference on Recommender systems. Chicago: ACM, 2011: 325−328.

[16] LIU Jiahui, DOLAN P, PEDERSEN E R. Personalized news recommendation based on click behavior[C]//Proceedings of the 15th international conference on Intelligent user interfaces. Hong Kong: ACM, 2010: 31−40.

[17] DAI Sunhao, SHAO Ninglu, ZHAO Haiyuan, et al. Uncovering ChatGPT's capabilities in recommender systems[EB/OL]. (2023−05−03)[2023−08−25]. http://arxiv.org/abs/2305.02182.

[18] WANG Lei, ZHANG Jingsen, CHEN Xu, et al. User behavior simulation with large language model based agents[EB/OL]. (2023−06−05)[2023−08−25]. https://arxiv.org/abs/2306.02552.

[19] QIU Zhaopeng, WU Xian, GAO Jingyue, et al. U-BERT: pre-training user representations for improved recommendation[J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(5): 4320−4327.

[20] SUN Fei, LIU Jun, WU Jian, et al. BERT4Rec: sequential recommendation with bidirectional encoder representations from transformer[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing: ACM, 2019: 1441−1450.

[21] COLIN R, NOAM S, ADAM R, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(1): 5485−5551.

[22] LI Jiacheng, WANG Ming, LI Jin, et al. Text is all you need: learning language representations for sequential recommendation[EB/OL]. (2023−05−23)[2023−08−25]. http://arxiv.org/abs/2305.13731.

[23] CUI Zeyu, MA Jianxin, ZHOU Chang, et al. M6-rec: generative pretrained language models are open-ended recommender systems[EB/OL]. (2022−05−17) [2023−08−25]. http://arxiv.org/abs/2205.08084.

[24] KOREN Y, BELL R, VOLINSKY C. Matrix factorization techniques for recommender systems[J]. Computer, 2009, 42(8): 30−37.

[25] ZHOU Guorui, ZHU Xiaoqiang, SONG Chenru, et al. Deep interest network for click-through rate prediction[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. London: ACM, 2018: 1059−1068.

[26] FRIEDMAN L, AHUJA S, ALLEN D, et al. Leveraging large language models in conversational recommender systems[EB/OL]. (2023−05−13)[2023−08−25]. http://arxiv.org/abs/2305.07961.

[27] LIN Guo, ZHANG Yongfeng. Sparks of artificial general recommender (AGR): early experiments with ChatGPT [EB/OL]. (2017−05−08)[2023−08−25]. http://arxiv.org/abs/2305.04518.

[28] XI Yunjia, LIU Weiwen, LIN Jianghao, et al. Towards open-world recommendation with knowledge augmenta-

tion from large language models[EB/OL]. (2023−06−19)[2023−08−25]. http://arxiv.org/abs/2306.10933.

[29] LI Jinming, ZHANG Wentao, WANG Tian, et al. GPT4Rec: a generative framework for personalized recommendation and user interests interpretation[EB/OL]. (2023−04−08)[2023−04−25]. http://arxiv.org/abs/2304.03879.

[30] DU Yingpeng, LUO Di, YAN Rui, et al. Enhancing job recommendation through LLM-based generative adversarial networks[EB/OL]. (2023−07−20)[2023−08−25]. http://arxiv.org/abs/2307.10747.

[31] WANG Wenjie, LIN Xinyu, FENG Fuli, et al. Generative recommendation: towards next-generation recommender paradigm[EB/OL]. (2023−04−07)[2023−08−25]. http://arxiv.org/abs/2304.03516.

[32] GENG Shijie, LIU Shuchang, FU Zuohui, et al. Recommendation as language processing (RLP): a unified pretrain, personalized prompt & predict paradigm (P5)[C]// Proceedings of the 16th ACM Conference on Recommender Systems. Seattle: ACM, 2022: 299−315.

[33] JI Jianchao, LI Zelong, XU Shuyuan, et al. GenRec: large language model for Generative recommendation[M]//Lecture Notes in Computer Science. Cham: Springer Nature Switzerland, 2024: 494−502.

[34] ZHANG Zizhuo, WANG Bang. Prompt learning for news recommendation[EB/OL]. (2023−04−11)[2023−08−25]. http://arxiv.org/abs/2304.05263.

[35] OKAMOTO K, CHEN Wei, LI Xiangyang. Ranking of closeness centrality for large-scale social networks[C]// PREPARATA FP, WU X, YIN J. International Workshop on Frontiers in Algorithmics. Berlin: Springer, 2008: 186−195.

[36] ZHANG Junlong, LUO Yu. Degree centrality, betweenness centrality, and closeness centrality in social network [C]//Proceedings of the 2017 2nd International Conference on Modelling, Simulation and Applied Mathematics. Paris: Atlantis Press, 2017: 300−303.

[37] NEWMAN M E J. A measure of betweenness centrality based on random walks[J]. Social networks, 2005, 27(1): 39−54.

[38] HUANG Xiao, ZHANG Jingyuan, LI Dingcheng, et al. Knowledge graph embedding based question answering[C]// Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining. Melbourne: ACM, 2019: 105−113.

[39] ZHANG Yuyu, DAI Hanjun, KOZAREVA Z, et al. Variational reasoning for question answering with knowledge graph[EB/OL]. (2017−11−27)[2023−01−01]. http://arxiv.org/abs/1709.04071.

[40] RONG Yu, HUANG Wenbing, XU Tingyang, et al. DropEdge: towards deep graph convolutional networks on node classification[EB/OL]. (2019−07−25)[2023−08−25]. http://arxiv.org/abs/1907.10903.

[41] ERRICA F, PODDA M, BACCIU D, et al. A fair comparison of graph neural networks for graph classification [EB/OL]. (2019−12−20)[2023−08−25]. http://arxiv.org/abs/1912.09893.

[42] WANG Heng, FENG Shangbin, HE Tianxing, et al. Can language models solve graph problems in natural language?[EB/OL]. (2023−05−17)[2023−08−25]. http://arxiv.org/abs/2305.10037.

[43] XIE Han, ZHENG Da, MA Jun, et al. Graph-aware language model pre-training on a large graph corpus can help multiple graph applications [EB/OL]. (2023−06−05)[2023−08−25]. http://arxiv.org/abs/2306.02592.

[44] GUO Jiayan, DU Lun, LIU Hengyu, et al. GPT4Graph: can large language models understand graph structured data? an empirical evaluation and benchmarking [EB/OL]. (2023−05−24) [2023−08−25]. http://arxiv.org/abs/2305.15066.

[45] WU Likang, QIU Zhaopeng, ZHENG Zhi, et al. Exploring large language model for graph data understanding in online job recommendations[EB/OL]. (2023−07−10)[2023−08−25]. http://arxiv.org/abs/2307.05722.

[46] ZHU Deyao, CHEN Jun, SHEN Xiaoqian, et al. MiniGPT-4: enhancing vision-language understanding with advanced large language models [EB/OL]. (2023−04−20)[2023−08−25]. http://arxiv.org/abs/2304.10592.

[47] LI Junnan, LI Dongxu, SAVARESE S, et al. BLIP−2: bootstrapping language-image pre-training with frozen image encoders and large language models [EB/OL]. (2023−01−30) [2023−08−25]. http://arxiv.org/abs/2301.12597.

[48] BAO Keqin, ZHANG Jizhi, ZHANG Yang, et al. TALLRec: an effective and efficient tuning framework to align large language model with recommendation [EB/OL]. (2023−04−30) [2023−08−25]. http://arxiv.org/abs/2305.00447.

[49] ZHANG Junjie, XIE Ruobing, HOU Yupeng, et al. Recommendation as instruction following: a large language model empowered recommendation approach [EB/OL]. (2023−05−11) [2023−08−25]. http://arxiv.org/abs/2305.07001.

[50] KANG Wangcheng, NI Jianmo, MEHTA N, et al. Do LLMs understand user preferences? evaluating LLMs on user rating prediction [EB/OL]. (2023−05−10) [2023−

08−25]. http://arxiv.org/abs/2305.06474.

[51] YONA G, HONOVICH O, LAISH I, et al. Surfacing biases in large language models using contrastive input decoding[EB/OL]. (2023−05−12)[2023−08−25]. http://arxiv.org/abs/2305.07378.

[52] CHEN Zhipeng, ZHOU Kun, ZHANG Beichen, et al. ChatCoT: tool-augmented chain-of-thought reasoning on chat-based large language models[EB/OL]. (2023−05−23) [2023−08−25]. http://arxiv.org/abs/2305.14323.

[53] OUYANG Long, WU J, XU Jiang, et al. Training language models to follow instructions with human feedback [EB/OL]. (2022−03−04)[2023−08−25]. http://arxiv.org/abs/2203.02155.

[54] HUA Wenyue, GE Yingqiang, XU Shuyuan, et al. UP5: unbiased foundation model for fairness-aware recommendation[EB/OL]. (2023−05−20)[2023−08−25]. http://arxiv.org/abs/2305.12090.

[55] ZHANG Jizhi, BAO Keqin, ZHANG Yang, et al. Is ChatGPT fair for recommendation? evaluating fairness in large language model recommendation[EB/OL]. (2023−05−12) [2023−08−25]. http://arxiv.org/abs/2305.07609.

[56] AI Qingyao, BAI Ting, CAO Zhao, et al. Information retrieval meets large language models: a strategic report from chinese IR community[J]. AI open, 2023, 4: 80−90.

[57] ZHAO W X, ZHOU Kun, LI Junyi, et al. A survey of large language models[EB/OL]. (2023−03−21)[2023−08−25]. http://arxiv.org/abs/2303.18223.

[58] BILLS S, CAMMARATA N, MOSSING D, et al. Language models can explain neurons in language models [EB/OL]. (2023−05−09)[2024−03−25]. https://openaipublic.blob. core. windows. net/neuron-explainer/paper/index.

[59] WEI Wei, REN Xubin, TANG Jiabin, et al. LLMRec: large language models with graph augmentation for recommendation[EB/OL]. (2023−11−01)[2024−03−25]. http://arxiv.org/abs/2311.00423.

作者简介:

吴国栋, 副教授, 主要研究方向为深度学习、推荐系统。主持安徽省自然科学研究重点项目 1 项、一般项目 1 项、安徽省科技攻关重点项目 1 项。发表学术论文 30 余篇。E-mail: wugd@ahau.edu.cn。

秦辉, 硕士研究生, 主要研究方向为推荐系统。E-mail: 2504864202@qq.com。

胡全兴, 硕士研究生, 主要研究方向区块链可信推荐。E-mail: 1763273299@qq.ocm。