

用于对话推荐系统的知识图谱检索增强生成

王本帅, 沙立伟, 邹丽, 侯铁

(山东建筑大学 计算机科学与技术学院, 山东 济南 250101)

摘要: 随着互联网的高速发展, 人们已经处于信息过载的时代, 通过对话理解用户动态偏好并为其推荐个性化信息具有重要意义。针对大语言模型(LLM)在对话推荐任务中存在的领域知识匮乏和幻觉问题, 提出一种知识图谱检索增强生成的创新方法。该方法的核心在于将知识图谱转化为自然语言文档, 并根据用户—系统对话动态检索相关知识, 将其整合到预定义的模板中以生成精准推荐, 不仅对齐了知识图谱与LLM的语义空间, 还为LLM提供了丰富的领域知识, 显著缓解了幻觉问题。此外, 该方法通过将推荐结果融入响应生成提示模板, 利用LLM生成对话, 实现了对话模块与推荐模块的一致性。实验结果表明, 该方法在两个公开数据集上的推荐准确性和对话响应质量均优于现有方法。消融实验也进一步验证了该方法中每个组件的有效性。

关键词: 对话推荐系统; 知识图谱; 大语言模型; 检索增强生成; 自然语言处理

DOI: 10.11907/tj.251186

扫描二维码阅读全文:



中图分类号: TP391.1; TP18

文献标识码: A

文章编号: 1672-7800(2026)007-0134-06

Knowledge Graph Retrieval-Augmented Generation for Conversational Recommender Systems

WANG Benshuai, SHA Liwei, ZOU Li, HOU Tie

(School of Computer Science, Shandong Jianzhu University, Ji'nan 250101, China)

Abstract: With the rapid development of the internet, people are in an era of information overload. Understanding users' dynamic preferences through conversational and providing personalized recommendations is of great significance. To address the issues of specific knowledge scarcity and hallucination in conversational recommendation systems using large language models (LLM), this study proposes an innovative method based on Knowledge Graph Retrieval-Augmented Generation. The core of this method lies in transforming knowledge graphs into natural language documents and dynamically retrieving relevant knowledge based on user-system dialogues, which is then integrated into predefined templates to generate accurate recommendations. This approach effectively aligns the semantic spaces of knowledge graphs and LLM and enriches LLM with extensive domain knowledge, thereby significantly mitigating hallucination issues. Furthermore, by seamlessly incorporating recommendation results into LLM, the method achieves coordinated responses between the dialogue module and the recommendation module. Experimental results demonstrate that the proposed method outperforms existing approaches in terms of recommendation accuracy and dialogue response quality on two public datasets. Ablation studies further validate the effectiveness of each component in the method.

Key Words: conversational recommender systems; knowledge graph; large language models; retrieval-augmented generation; natural language processing

0 引言

随着互联网的飞速发展, 信息过载已成为当今社会的

显著特征。面对海量信息, 推荐系统可以通过分析用户行为数据, 构建用户画像并推测其偏好与意图, 从而为用户提供个性化推荐。传统推荐算法主要通过分析用户的隐式反馈(如点击、查询记录等)以评估用户偏好^[1]。然而,

收稿日期: 2025-06-04

基金项目: 国家自然科学基金项目(62406051)

作者简介: 王本帅(1997-), 男, CCF 学生会会员, 山东建筑大学计算机科学与技术学院硕士研究生, 研究方向为自然语言处理、对话推荐系统、大语言模型; 沙立伟(2000-), 男, 山东建筑大学计算机科学与技术学院硕士研究生, 研究方向为智能信息处理; 邹丽(1971-), 女, 博士, 山东建筑大学计算机科学与技术学院教授、硕士生导师, 研究方向为模糊信息处理、非经典逻辑、不确定性推理; 侯铁(1978-), 女, 博士, 山东建筑大学计算机科学与技术学院讲师, 研究方向为数理逻辑、软件形式化、本体工程、数据分析。本文通讯作者: 侯铁。

用户偏好具有动态性,新用户和新物品缺乏历史交互数据,导致传统方法难以实时捕捉用户偏好,从而产生推荐偏差。近年来,自然语言处理的快速发展,尤其是大语言模型(Large Language Models, LLM)这一革命性技术的出现,为众多领域提供了新的研究思路。基于自然语言处理技术的推荐方法受到广泛关注,其中对话推荐系统(Con-

versational Recommender Systems, CRS)成为一个重要的研究方向^[2]。

如表 1 所示(其中提到的电影和实体分别用斜体和加粗标注),CRS 通过与用户进行多轮对话,深入理解用户需求,推断用户实时偏好,最终生成高质量的推荐和流畅的对话回复。

Table 1 An illustrative example of a conversation on movie recommendation between a user and the system

表 1 一个关于电影推荐的用户与系统对话的示例

角色	对话内容
User:	Hey! I'm looking for movie like <i>The Sixth Sense</i> , can you help?
System:	How about <i>The Happening</i> ? Or have you seen <i>Shutter Island</i> ?
User:	Yes I loved <i>Shutter Island</i> but haven't seen <i>The Happening</i> .
System:	Along the lines of <i>The Happening</i> , there is also <i>Signs</i> directed by M. Night Shyamalan .
User:	Oh good point I did love <i>Signs</i> (2008). Thank you! Bye!
System:	Bye!

因此,CRS 的核心任务可归纳为两个方面:①推荐,即基于对话从候选项目集中筛选出符合用户需求的项目;②对话,即生成包含推荐结果的对话并返回给用户。

近年来,LLM 作为一项突破性技术,展现出巨大的潜力,并在医学、法律等多个领域展现了其卓越性能^[3-4]。传统推荐系统也已成功通过融合 LLM 分析用户交互历史,从而生成高质量的推荐结果^[5]。然而,目前仅有少数研究探索了如何利用 LLM 理解 CRS 的上下文信息,He 等^[6]研究了设计提示模板,利用 LLM 的零样本推理能力生成最终推荐。但由于 LLM 在特定领域知识上的局限性以及固有的幻觉性问题,其生成的回复往往看似合理实则错误。研究表明,通过为 LLM 注入相关知识,可以有效缓解其幻觉问题,从而提升推荐系统的准确性和可靠性^[7]。

基于上述背景,本文提出了一种基于知识图谱检索增强生成的对话推荐系统方法(Knowledge Graph Retrieval-Augmented Generation Conversational Recommender Systems, KRCRS)。在推荐子任务中,KRCRS 利用对话中出现的实体关系增强知识图谱(Knowledge Graph, KG),设计转换规则将 KG 转化为自然语言文档,并从中检索与用户对话上下文相关的知识,增强预定义的推荐提示模板,随后利用 LLM 的零样本推理能力生成推荐结果。这一机制为 LLM 融入领域知识的同时缓解了幻觉性问题。在对话子任务中,KRCRS 将对话历史与推荐结果整合到响应生成提示模板中,充分发挥 LLM 的零样本推理能力,确保生成的回复内容与推荐结果相一致。本文贡献点如下:①提出了一种 KRCRS 的创新性模型,通过将检索增强生成技术与 LLM 相结合,有效缓解了 LLM 的幻觉问题,显著提升了推荐结果的准确性和可靠性;②设计了一种高效的 KG 转换规则,将 KG 转化为自然语言文档,为 LLM 提供了丰富的领域知识,从而增强了模型对上下文的理解能力和推理能力;③通过大量实验验证了 KRCRS 在多个数据集上的优越性能,其在推荐子任务和对话子任务中均取得了显著提升。此外,消融实验进一步证实了 KRCRS 框架中关键组件的有效性。

1 相关工作

本文对模型所涉及的两个核心研究方向进行了综述:①系统回顾了 CRS 的相关研究进展;②全面梳理了 LLM 在推荐系统领域的应用现状与发展趋势。

1.1 对话推荐系统

CRS 因其显著的实际应用价值,已成为推荐系统领域的重要研究方向。现有 CRS 主要可分为两类:基于属性的 CRS 和生成式 CRS。

基于属性的 CRS 主要聚焦于优化推荐策略。这类方法通常通过向用户询问特定属性相关问题以实现信息过滤^[2],其核心在于询问哪些属性、是否继续提问以及如何生成最终推荐。这类策略的优势在于能够以最少的用户交互实现精准推荐。部分研究采用多臂老虎机和强化学习等技术以优化决策过程^[2,8]。然而,这类方法往往难以生成高质量、拟人化的对话响应,且过度依赖预定义模板,限制了系统的灵活性,影响用户体验。

生成式 CRS 则致力于同时实现精准推荐和自然对话生成。与基于属性的方法不同,这类方法通常整合了对话系统模块。例如,Li 等^[9]将自动编码器与循环神经网络相结合用于对话生成。部分研究还尝试引入外部知识,如 KG^[10]。然而,KG 实体与对话文本之间的语义鸿沟为二者的有效融合带来了挑战。针对这一问题,Zhou 等^[11]提出基于对比学习的外部信息融合方法,Wang 等^[12]开发了基于预训练语言模型的知识增强提示学习框架。尽管这些研究取得了一定进展,但仍缺乏对文本深层次语义的理解能力。本文提出的 KRCRS 利用 LLM 生成推荐,可以有效提升上下文理解能力。

1.2 大语言模型推荐系统

LLM 在众多下游任务中展现了卓越的性能,其中也包括推荐系统领域^[13-14]。早期的研究方法主要利用 LLM 的零样本推理能力,将用户行为转化为文本表示,并通过提

示模板生成推荐^[15]。然而,由于这些方法缺乏特定领域的知识,导致推荐准确性受到限制。为此,文献[5]通过对用户行为序列进行微调以增强 LLM 的性能,但这种方法不仅消耗大量计算资源,还可能导致生成的项目超出候选集范围,从而影响推荐精准性。文献[16]引入了项目 ID 或索引机制,但这种方式会破坏语义连贯性,进而引发幻觉性问题。本文提出了一种创新的解决方案,通过将 KG 转化为自然语言文档,并利用检索器动态检索相关知识以增强 LLM。这种方法不仅显著提升了推荐效率,还有效缓解了 LLM 的幻觉问题,从而实现了更精准和可靠的推荐。

2 知识图谱检索增强生成模型

2.1 问题定义

本文提出的 KRCRS 模型通过与用户对话分析用户偏好,并以对话形式向用户反馈推荐结果。与传统 CRS 模型类似,KRCRS 同样包含两个核心模块:推荐模块和对话模块。

设 $C = \{s_t\}_{t=1}^n$ 表示用户和系统对话历史,其中 s_t 表示第 t 轮对话,并且每一轮对话 $s_t = \{w_j\}_{j=1}^m$, w_j 表示来自词汇表 V 表示一句话中的第 j 个单词,在第 t 轮,给定对话历史 C ,推荐模块从项目集 I 中挑选出项目 I_t 推荐给用户,对话模块生成包含项目 I_t 的回复 $R = s_{t+1}$ 。

2.2 整体架构

KRCRS 整体架构如图 1 所示。

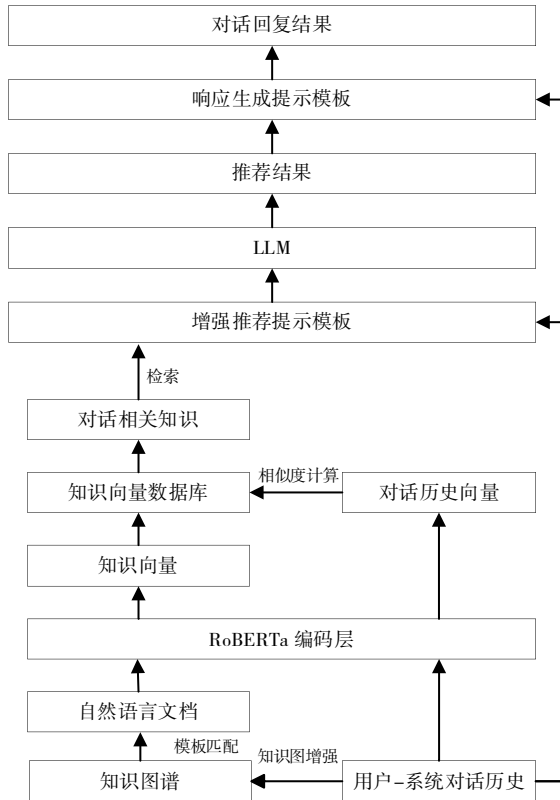


Fig. 1 The overall architecture of KRCRS

图1 KRCRS 整体架构

首先,通过抽取对话中实体关系增强外部 KG,使得 KG 具备更丰富的领域知识,并利用规则转换将增强后的外部 KG 转化为自然语言文档,这一步骤旨在使 KG 与 LLM 的语义空间对齐;其次,利用嵌入模型将自然语言文档编码得到知识向量并进行存储;再次,将用户与系统的对话内容同样转换为向量,并从自然语言文档中检索对话相关知识;接着,结合检索结果增强提示模板,引导 LLM 生成最终推荐结果;最后,将推荐结果整合到预定义的响应生成模板中,生成包含推荐信息的自然语言回复。

2.3 知识图谱文档库构建

KG 包含的数据通常是事实性的,信息也通常从可信的语料库中提取。本文选择 DBpedia 作为外部 KG,并从 DBpedia 中抽取包含相关实体的三元组以生成子图 $\mathcal{G}_e, \mathcal{G}_r$ 以 $[e_1, r, e_2]$ 的三元组形式出现^[17]。其中, e_1 表示三元组头部实体, e_2 表示三元组尾部实体, r 表示实体 e_1 和 e_2 之间的关系。由于 DBpedia 是一个广泛使用的 KG,缺乏较详细的领域知识,因此为了增强 KG,本文通过抽取数据集中同一个对话共同出现的实体建立连接,并添加新的关系“Co-occurrence”。为了将三元组转化为自然语言,本文为子图 $\mathcal{G}_e$ 设置了固定的自然语言模板,图 2 展示了自然语言模板。通过深度优先遍历算法从 $\mathcal{G}_e$ 中提取所有三元组依次填入模板进行填充从而生成自然语言句子。

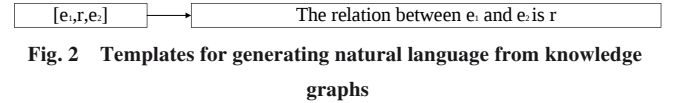


Fig. 2 Templates for generating natural language from knowledge graphs

图2 KG 中生成自然语言的模板

2.4 知识存储与检索

由于 KG 涉及较少的上下文关系以及主题关系,本文直接将每个句子作为一个独立的分块单元,并利用 RoBERTa 编码器将这些句子编码为向量^[18]。RoBERTa 采用了 24 层 Transformer 编码器结构,向量维度为 1 024,并通过大规模语料进行无监督预训练,能够有效捕捉文本的上下文语义信息。随后,本文使用 FAISS^[19] 构建向量数据库以存储这些向量。由于知识库的数据量庞大,无法将所有知识直接用于增强生成过程,因此通过检索筛选出最相关的信息,可以在减少噪声的同时提升生成效率。FAISS 通过其优化的索引结构和算法,能够在大规模数据集上实现高效的相似性搜索。

当需要根据用户对话历史进行推荐时,本文首先利用 RoBERTa 将用户与系统的对话历史编码为向量,再计算其与向量数据库中分块单元的向量相似度,筛选出最相似的 N 个片段。而由于知识库规模的庞大,直接计算所有分块单元相似度效率较低。为此,在进行相似度计算之前,本文对知识库进行分析,抽取与对话相关实体的三阶关系并将其作为候选知识,随后与候选知识分块单元的向量进行相似度计算,这一步骤极大缩减了检索时间。本文选用余弦相似度作为相似度计算方法,主要原因在于对话和文档的长度差异较大,而余弦相似度更注重向量方向上的相

似性,能够在一定程度上缓解文档长度差异所带来的偏差。此外,余弦相似度在高维向量相似度计算中表现更为稳定和高效。余弦相似度计算公式如式(1)所示。

$$\text{余弦相似度}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (1)$$

其中, A, B 表示两个向量, $\cdot$ 表示向量的点积, $\|A\| \|B\|$ 分别表示向量的模长。通过这种方式,本文能够高效地从大规模知识库中检索出与用户对话历史最相关的信息,从而为生成过程提供高质量的上下文支持。

2.5 提示增强输出

通过知识检索,可以获取与当前对话相关的外部知识。这些知识将与推荐生成模板相结合,模板主要包括以下3部分内容:指令、用户与系统的对话上下文以及检索到的对话相关知识。增强推荐提示模板如图3所示。

Instruction:
Here's a conversation about movies between user and recommender system. Pretend you're the recommender system and I'll also give some related knowledge about conversation, please recommend a movie to user according to these information.

Conversation:
{用户-系统对话历史}

Related Knowledge:
{对话相关知识}

Fig. 3 Prompt template for augmented recommendation

图3 增强推荐提示模板

其中,指令用于明确指导LLM当前的任务目标,用户与系统的对话上下文提供了用户的偏好和历史交互信息,而外部知识则为LLM提供了更丰富的背景信息和参考依据。在该模块中,本文采用LLaMA2-7b作为LLM,原因在于其开源特性可以有效保证用户隐私安全且在多种任务中表现出优异性能^[20]。此外,LLaMA2-7b支持长达4 096的上下文长度,这使得推荐模板能够容纳更多的外部知识,从而提升生成结果的质量。

2.6 响应生成模块

通过知识增强后,LLM能够生成可靠的推荐结果。然而,由于CRS与用户的交互是以对话形式进行,如果直接展示推荐结果,可能会降低用户体验。因此,将推荐结果自然地融入对话中显得尤为重要。为了生成最终的对话式回复,本文设计了一个响应生成模块。该模块利用LLM的零样本推理能力,能够生成流畅且包含推荐结果的回复。具体而言,在第 $t+1$ 轮,本文设计了一个响应生成提示模板 P_{res} ,然后将用户-系统对话历史 C ,推荐结果 I_t 融入到模板中,无需对参数进行调整。响应生成提示模板如图4所示。

LLM根据融入后的提示生成包含推荐结果 I_t 的回复 R ,如式(2)所示。

$$R = LLM(P_{res}, C, I_t) \quad (2)$$

其中,LLM表示Llama2-7b模型。通过这种方式,系统能够生成既符合用户需求又自然流畅的对话回复,并且保

Instruction:
Here's a conversation about movies between user and a movie recommender system. Pretend you're a movie recommender system and I'll also give a movie you will recommend to user, Please generate an appropriate response.

Conversation:
{用户-系统对话历史}

movie:
{推荐结果}

Fig. 4 Prompt template for response generation

图4 响应生成提示模板

证了推荐和对话的一致性,提升了用户交互体验。

3 实验与结果分析

3.1 数据集

本文在两个公开数据集ReDial^[9]和INSPIRED^[21]上进行了实验,这两个数据集被广泛用于CRS的评估,专注于电影推荐场景。表2详细列出了所用数据集的基本统计信息。其中,INSPIRED是一个小规模的英语CRS数据集,而ReDial则是一个大规模的英语CRS数据集,包含了更丰富的电影推荐相关对话和项目信息。通过在这两个数据集上的实验,能够全面评估模型在不同数据规模和应用场景下的性能表现。

Table 2 The statistics of datasets

表2 数据集统计

Datasets	#Dialogs	#Utterance	#Items
INSPIRED	1 001	35 811	1 783
ReDial	10 006	182 150	51 699

3.2 对比实验

在本文KRCRS中,有推荐和对话两个子任务,为了确保比较的全面性,对比方法不仅包括CRS方法,还包括传统的推荐方法和对话方法。

(1)Popularity。该方法分析语料库中的历史推荐,根据项目出现频率对其进行排序,并提供相应的推荐。

(2)TextCNN^[22]。该方法基于卷积神经网络模型,通过从对话上下文中提取用户特征对项目进行排序。

(3)Transformer^[23]。该模型使用一个六层的的编码器—解码器框架以生成对话响应。

(4)LLaMA2-7b。该方法是一个建立在Transformer架构上的预训练过的开源LLM,在开源LLM中效果优异。

(5)GPT-3.5-turbo^[24]。该模型是由OpenAI开发的闭源LLM,拥有强大的零样本学习能力,在各类任务上表现出优异的效果。

(6)ReDial^[9]。该模型包含一个基于HRED的对话框生成模块和一个基于自动编码器的推荐模块^[26]。

(7)KBRD^[10]。该方法利用DBpedia以增强上下文的语义表示。它使用了一个基于Transformer的对话生成模块,并在生成过程中结合了知识图信息以优化对话回复。

(8)KGSF^[25]。该方法结合了DBpedia和Conceptnet,

改进了项目和单词的语义表示,并使用互信息最大化以对齐不同组件的语义空间^[27]。

3.3 评估方法

本文使用一般的和标准的客观评价指标以验证该方法的性能。对于推荐任务,本文使用 Recall@K ($K=1, 10, 50$)作为评价指标,表示真正的目标项目是否包含在模型预测的前K个项目中。对于对话任务,本文使用 Distinct- n ($n=2, 3, 4$)作为评估指标,它计算 N-Gram 与生成文本中 N-Gram 总数的比率,其中 N-Gram 指文本中连续的 n 个词。

3.4 推荐结果评估

在推荐任务上的实验结果比较如表3所示,可以得到以下结论:①在传统的推荐方法中,TextCNN通过上下文内容为推荐提供更多的信息线索,其表现优于 Popularity,从而获得更高的准确性^[22];②CRS推荐方法准确率明显优于传统推荐方法,因为CRS通过与用户进行交互,更能理解用户的真实意图;③在CRS模型中,KBRD通过合并外部知识图以增强用户表示能力^[10];KGSF合并外部知识图的同时通过互信息最大化以增强实体表示,取得了更优结果^[25];④本文KRCRS显著优于所有的对比实验方法。在于它充分利用了LLM的理解能力,并通过将KG转化为自然语言对齐了和LLM的语义空间并提供外部知识,缓解了幻觉性问题。

Table 3 Comparison results of different methods on the recommendation task

表3 不同方法在推荐任务上的结果比较

Methods	Redial			INSPIRED		
	Recall@1	Recall@		Recall@1	Recall@	
		10	50		10	50
Popularity	0.012	0.061	0.179	0.032	0.155	0.323
TextCNN ^[22]	0.013	0.068	0.191	0.047	0.161	0.327
Redial ^[9]	0.024	0.140	0.320	0.030	0.117	0.285
KBRD ^[10]	0.031	0.150	0.336	0.058	0.172	0.265
KGSF ^[25]	0.038	0.183	0.378	0.064	0.175	0.273
OUR KRCRS	0.044	0.205	0.421	0.078	0.198	0.291

3.5 对话结果评估

表4展示了使用 Distinct- n ($n=2, 3, 4$)作为评估指标对对话进行评估的结果。可以得到以下结论:①与CRS方法相比,预训练的LLM可以生成高质量的回复,这要归功于LLM优秀的文本理解和生成能力,故GPT-3.5-turbo^[24]和LLaMA2-7b^[20]在Distinct指标方面具有更好的性能;②基于预训练LLM的方法中,GPT-3.5-turbo^[24]由于具有更多的参数,因此优于LLaMA2-7b^[20],这使其具有更强的对话生成能力;③在CRS方法中,KBRD通过利用基于知识图的单词偏差提高了低频单词的概率,取得了比ReDial更好的性能^[10];KGSF通过知识图增强的交叉注意层增强了实体和单词之间的交互^[25];④本文KRCRS优于所有对比方法,因为它将推荐结果集成到模板中,充分利用了LLM的生成能力,从而提高了生成响应的丰富性和多样性。

Table 4 Comparison of different methods on the conversation task

表4 不同方法在对话任务上结果比较

Methods	Redial			INSPIRED		
	Distinct- n			Distinct- n		
	2	3	4	2	3	4
Transformers ^[23]	0.148	0.151	0.137	1.02	2.248	3.582
LLaMA2-7b ^[20]	1.452	2.218	4.376	4.669	6.832	8.115
GPT-3.5-turbo ^[24]	1.702	3.201	4.576	5.331	7.019	8.936
Redial ^[9]	0.225	0.236	0.228	1.347	1.521	3.445
KBRD ^[10]	0.263	0.368	0.423	1.369	2.259	3.592
KGSF ^[25]	0.330	0.417	0.521	1.608	2.719	4.929
OUR KRCRS	1.835	3.368	4.902	5.417	7.226	9.259

3.6 消融实验

本文通过消融实验证明了KRCRS中每个组件的有效性。KRCRS的核心是利用外部知识增强LLM以生成更可靠的推荐,主要评估模型在推荐任务上的性能,结果如表5所示。具体实验设置如下:

(1)KRCRS w/o K。移除了知识检索增强,只使用LLM进行推荐。

(2)KRCRS w/o T。移除了KG转化为自然语言的过程,使用原始KG用于增强生成。

Table 5 Ablation study on the recommendation task

表5 推荐任务上的消融实验

Methods	Redial			INSPIRED		
	Recall@1	Recall@		Recall@1	Recall@	
		10	50		10	50
KRCRS w/o K	0.033	0.165	0.377	0.060	0.169	0.266
KRCRS w/o T	0.041	0.193	0.410	0.074	0.188	0.284
Our KRCRS	0.044	0.205	0.421	0.078	0.198	0.291

基于表5结果可以得到以下结论:①当知识检索增强模块被移除后,推荐准确率在Recall@1、Recall@10和Recall@50上都有大幅下降,这表明仅仅依赖LLM的生成远远不够,因为知识增强可以为LLM提供更多的事实依据;②当KG文档转化模块被移除后,推荐准确率在Recall@1、Recall@10和Recall@50上都有所下降,这表明转化后的KG文档更有助于LLM理解知识。

4 结语

本文提出了一种KRCRS的方法,通过对话中实体关系增强KG,并将KG转化为KG文档,在进行推荐时检索相关知识用于增强LLM,进一步提升了推荐质量,并利用LLM强大的生成能力,生成流畅且富有信息性的回复。实验结果表明,本文KRCRS在推荐任务和对话任务上都具有良好性能。未来,可构建一个多模态CRS,通过更丰富的内容,提供更精确的推荐结果,并为用户提供更多样化和更丰富的回复。

参考文献:

- [1] HE X, LIAO L, ZHANG H, et al. Neural collaborative filtering[C]//Proceedings of the 26th International Conference on World Wide Web, 2017: 173–182.
- [2] CHRISTAKOPOULOU K, RADLINSKI F, HOFMANN K. Towards conversational recommender systems[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016: 815–824.
- [3] LI Y, LI Z, ZHANG K, et al. ChatDoctor: a medical chat model fine-tuned on a large language model Meta-AI(LLaMA) using medical domain knowledge[J]. Cureus, 2023, 15(6): e40895.
- [4] YIN H, WU Z H, LIU T T, et al. Mixture of expert large language model for legal case element recognition[J]. Journal of Frontiers of Computer Science and Technology, 2024, 18(12): 3260–3271.
尹华, 吴梓浩, 柳婷婷, 等. 法律案件要素识别混合专家大模型[J]. 计算机科学与探索, 2024, 18(12): 3260–3271.
- [5] BAO K, ZHANG J, ZHANG Y, et al. TALLRec: an effective and efficient tuning framework to align large language model with recommendation[C]//Proceedings of the 17th ACM Conference on Recommender Systems, 2023: 1007–1014.
- [6] HE Z, XIE Z, JHA R, et al. Large language models as zero-shot conversational recommenders[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023: 720–730.
- [7] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459–9474.
- [8] LEI W, HE X, MIAO Y, et al. Estimation-action-reflection: towards deep interaction between conversational and recommender systems[C]//Proceedings of the 13th International Conference on Web Search and Data Mining, 2020: 304–312.
- [9] LI R, EBRAHIMI K S, SCHULZ H, et al. Towards deep conversational recommendations[DB/OL]. <https://arxiv.org/abs/1812.07617>.
- [10] CHEN Q, LIN J, ZHANG Y, et al. Towards knowledge-based recommender dialog system[DB/OL]. <https://arxiv.org/abs/1908.05391>.
- [11] ZHOU Y, ZHOU K, ZHAO W X, et al. C2-CRS: coarse-to-fine contrastive learning for conversational recommender system[C]//Proceedings of the 15th ACM International Conference on Web Search and Data Mining, 2022: 1488–1496.
- [12] WANG X, ZHOU K, WEN J, et al. Towards unified conversational recommender systems via knowledge-enhanced prompt learning[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022: 1929–1937.
- [13] ZHANG Z W, WU Z X, ZHANG W. Intelligent practice exploration of large language model in medical emergency knowledge graph question-and-answer service[J]. Software Guide, 2024, 23(12): 143–149.
张子威, 武志学, 张薇. 大语言模型在医疗应急知识图谱问答服务中的智能化实践探索[J]. 软件导刊, 2024, 23(12): 143–149.
- [14] HUANG W, LI Z J. Recommendation systems: a comparative study of traditional models and pre-trained model approaches[J]. Software Guide, 2025, 24(2): 204–210.
黄文, 李振江. 传统模式与大语言模型下推荐系统的比较研究[J]. 软件导刊, 2025, 24(2): 204–210.
- [15] GAO Y, SHENG T, XIANG Y, et al. Chat-REC: towards interactive and explainable LLMs-augmented recommender system[DB/OL]. <https://arxiv.org/abs/2303.14524>.
- [16] HUA W, XU S, GE Y, et al. How to index item ids for recommendation foundation models[C]//Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region, 2023: 195–204.
- [17] BIZER C, LEHMANN J, KOBILAROV G, et al. DBpedia—a crystallization point for the web of data[J]. Journal of Web Semantics, 2009, 7: 154–165.
- [18] LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[DB/OL]. <https://arxiv.org/abs/2303.14524>.
- [19] DOUZE M, GUZHVA A, DENG C, et al. The faiss library[DB/OL]. <https://arxiv.org/abs/2303.14524>.
- [20] TOUVRON H, MARTIN L, STONE K, et al. LLaMA 2: open foundation and fine-tuned chat models[DB/OL]. <https://arxiv.org/abs/2303.14524>.
- [21] HAYATI S A, KANG D, ZHU Q, et al. Inspired: toward sociable recommendation dialog systems[DB/OL]. <https://arxiv.org/abs/2303.14524>.
- [22] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP), 2014: 1746–1751.
- [23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[DB/OL]. <https://arxiv.org/abs/1706.03762>.
- [24] SCHULMAN J, ZOPH B, KIM C, et al. ChatGPT: optimizing language models for dialogue[DB/OL]. <https://openai.com/blog/chatgpt>.
- [25] ZHOU K, ZHAO W X, BIAN S, et al. Improving conversational recommender systems via knowledge graph based semantic fusion[C]//Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020: 1006–1014.
- [26] SERBAN I V, SORDONI A, LOWE R, et al. A hierarchical latent variable encoder-decoder model for generating dialogues[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2017: 3295–3301.
- [27] SPEER R, CHIN J, HAVASI C. ConceptNet 5.5: an open multilingual graph of general knowledge[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2017: 4444–4451.

(责任编辑:孙娟)