https://crad.ict.ac.cn
62(3): 605−619, 2025

大语言模型和知识图谱协同的跨域异质数据查询框架

吴文隆[1] 尹海莲[4] 王 宁[1] 徐梦飞[1] 赵鑫喆[1] 殷崭祚[1] 刘元睿[1] 王昊奋[5] 丁 岩[6] 李博涵[1,2,3]

[1](南京航空航天大学人工智能学院 南京 211106)

[2](脑机智能技术教育部重点实验室(南京航空航天大学) 南京 211106)

[3](智能决策与数字化运营工业和信息化部重点实验室(南京航空航天大学) 南京 211106)

[4](中小型无人机先进技术工业和信息化部重点实验室(南京航空航天大学) 南京 210006)

[5](同济大学创意设计学院 上海 200092)

[6](中移(苏州)软件技术有限公司数据库产品部 南京 210004)

(wuwenlong@nuaa.edu.cn)

A Synergetic LLM-KG Framework for Cross-Domain Heterogeneous Data Query

Wu Wenlong[1], Yin Hailian[4], Wang Ning[1], Xu Mengfei[1], Zhao Xinzhe[1], Yin Zhanzuo[1], Liu Yuanrui[1], Wang Haofen[5], Ding Yan[6], and Li Bohan[1,2,3]

[1](*College of Artificial Intelligence*, *Nanjing University of Aeronautics and Astronautics*, *Nanjing* 211106)

[2](*Key Laboratory of Brain-Machine Intelligence Technology* (*Nanjing University of Aeronautics and Astronautics*), *Ministry of Education*, *Nanjing* 211106)

[3](*Key Laboratory of Intelligent Decision and Digital Operation* (*Nanjing University of Aeronautics and Astronautics*), *Ministry of Industry and Information Technology*, *Nanjing* 211106)

[4](*Key Laboratory of Advanced Technology for Small and Medium-sized UAVs* (*Nanjing University of Aeronautics and Astronautics*), *Ministry of Industry and Information Technology*, *Nanjing* 210006)

[5](*College of Design and Innovation*, *Tongji University*, *Shanghai* 200092)

[6](*China Mobile* (*Suzhou*) *Software Technology Co.Ltd.*, *Nanjing* 210004)

Abstract Recent advances in large language models (LLMs) have significantly elevated requirements for data quality in practical applications. Real-world scenarios often involve heterogeneous data from multiple correlated domains. Yet cross-domain data integration remains challenging due to privacy and security concerns that prohibit centralized sharing, thereby limiting LLM's effective utilization. To address this critical issue, we propose a novel framework integrating LLM with knowledge graphs (KGs) for cross-domain heterogeneous data query. Our approach presents a systematic governance solution under the LLM-KG paradigm. First, we employ domain adapters to fuse cross-domain heterogeneous data and construct corresponding KG. To enhance query efficiency, we introduce knowledge line graphs and develop a homogeneous knowledge graph extraction (HKGE) algorithm for graph reconstruction, significantly improving cross-domain data governance performance. Subsequently, we propose a trusted subgraph matching algorithm TrustHKGM to ensure high-confidence multi-domain queries through confidence computation and low-quality node filtering. Finally, we design a multi-domain knowledge line graph prompting (MKLGP) algorithm to enable efficient and trustworthy cross-domain query answering within the LLM-KG framework. Extensive experiments on multiple real-world datasets demonstrate the superior effectiveness and efficiency of our approach compared with state-of-the-art solutions.

收稿日期: 2024-07-21; **修回日期**: 2025-02-10

基金项目: "十四五"民用航天技术预先研究项目(D020101)

This work was supported by the Pre-research Project of Civil Space Technology during the 14th Five-Year Plan Period (D020101).

Key words large language model; cross-domain heterogeneous data; knowledge graph; multi-domain data query; data governance

摘 要 大语言模型(large language model, LLM)技术热潮对数据质量的要求提升到了一个新的高度. 在现实场景中, 数据通常来源不同且高度相关. 但由于数据隐私安全问题, 跨域异质数据往往不允许集中共享, 难以被 LLM 高效利用. 鉴于此, 提出了一种 LLM 和知识图谱(knowledge graph, KG)协同的跨域异质数据查询框架, 在 LLM+KG 的范式下给出跨域异质数据查询的一个治理方案. 为确保 LLM 能够适应多场景中的跨域异质数据, 首先采用适配器对跨域异质数据进行融合, 并构建相应的知识图谱. 为提高查询效率, 引入线性知识图, 并提出同源知识图抽取算法 HKGE 来实现知识图谱的重构, 可显著提高查询性能, 确保跨域异质数据治理的高效性. 进而, 为保证多域数据查询的高可信度, 提出可信候选子图匹配算法 TrustHKGM, 用于检验跨域同源数据的置信度计算和可信候选子图匹配, 剔除低质量节点. 最后, 提出基于线性知识图提示的多域数据查询算法 MKLGP, 实现 LLM+KG 范式下的高效可信跨域查询. 该方法在多个真实数据集上进行了广泛实验, 验证了所提方法的有效性和高效性.

关键词 大语言模型; 跨域异质数据; 知识图谱; 多域数据查询; 数据治理

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440634 **CSTR:** 32373.14.issn1000-1239.202440634

数据, 是继土地、劳动力、资本、技术四大生产要素之后的第五大生产要素, 我国已提出要加快培育数据要素市场[1]. 以大语言模型(large language model, LLM)为代表的生成式人工智能技术更需要高质量的数据和丰富的语义知识支持. 因此, 面向大语言模型的跨域数据共享和流通对于实现数据要素价值最大化变得至关重要[2].

跨域异质数据指来源于不同行业或领域, 具有不同数据模型、结构、语义和质量标准的数据集合, 其具有来源多样、质量参差不齐、管理分散、形式异构的特点[3-7]. 虽然目前针对跨域异质数据的研究已经取得诸多进展, 但是大多数工作都依赖于专家引擎和规则系统[8-9], 或将任务建模为语义匹配问题或语义匹配任务[10-12]. 这些方法存在局限性, 它们通常需要大量的标注数据或提供足够多的专家知识. 此外, 面向千行百业多样化的数据规范标准, 多模态、多层次的数据格式, 高品质、海量化的数据需求, 数据治理的范围已经大大扩展[13]. 尤其对于训练行业大语言模型而言, 处理各种模态的数据变得至关重要. 而在需要实时数据调整的数据科学场景中, 跨域异质数据之间存在复杂的依赖关系, 这会对 LLM 的推理性能带来一定的损失[14]. 再者, 数据的安全隐私保护和规范化也是必须要考虑的问题. 这些问题严重阻碍了数据价值潜力的充分释放[15].

大语言模型的兴起和广泛应用为数据治理带来新的挑战与契机. 凭借其先进的自然语言处理能力, LLM 能够从多个异质数据源中提取有价值的见解, 如需求文本文档和仿真模型文件. 这反过来将推动更有效的数据使用, 从而导致更好的决策和业务结果.

知识图谱(knowledge graphs, KG)作为大数据时代的知识工程集大成者, 可以为跨域异质数据的知识化组织和智能应用提供有效的解决方案[16]. 知识图谱可以进行复杂、非结构化的跨域多模态数据组织, 从而生成准确、一致地描述的高质量结构化数据图[17]. 同时可以有效缓解黑盒大语言模型存在的幻觉问题、灾难性遗忘以及错误的知识编辑等问题, 有效赋能 LLM 时代数据治理的数据规范性和可解释性, 拓宽应用场景, 提高治理效率[18-22].

据此, 本文结合了 LLM 与 KG 的优势, 提出一种 LLM+KG 协同的跨域异质数据处理框架. 通过 KG 的高效图结构对跨域异质数据进行数据融合和结构化知识体系的构建, 从而利用 KG 为 LLM 提供精准、可靠的领域知识, 增强 LLM 回答的准确性、忠实性和可解释性. 同时利用 LLM 强大的语义理解和信息抽取能力, 助力知识图谱的动态构建、补全, 保障 KG 中跨领域数据的一致性. 最终, 本文在 4 个公开数据集上进行了跨域冲突数据的置信度计算, 以及多域数据查询的性能进行大量实验分析, 证明了所提方法的合理性.

综上所述, 本文总结 3 点贡献:

1)利用适配器结构, 实现大规模的跨域异质数据整合和 KG 构建. 提出同源知识图抽取算法, 引入线性知识图重构 KG, 保证大规模跨域数据治理的高效性.

2)在数据查询方面, 提出可信候选子图匹配算

法，用于进行同源数据的一致性检验，同时剔除低质量节点，保证多域数据查询结果的高可信性.

3）提出基于线性知识图提示的多域数据查询算法，用于 LLM+KG 的高效可信跨域数据查询. 同时，针对跨域异质数据的置信度计算和数据查询 2 个模块，在 4 个公开数据集上进行了广泛全面的实验，证明了所提方法的有效性和高效性.

1 相关工作

1.1 KG 用于数据治理

KG 能够将分散的数据整合到一个统一的语义化知识体系中，通过实体、属性、关系等语义元素的建模，更好地发现数据之间的内在联系. 因此，如何从多源异构数据出发，构建高质量的知识图谱，从而更好地赋能大语言模型已经成为当前大语言模型和大数据领域的热点研究方向.

现有工作面向特定领域大规模的数据和知识融合过程，针对融合算法的效率、多源数据的数据质量评估、基于语义的数据和知识融合等方面开展研究. 杨晓慧等人[23]提出一种分布式的 KG 表示学习算法，广泛用于知识图谱链路预测和多标签分类，为跨域异质数据的治理提供了一个有效的解决途径. 董永强等人[24]面向复杂且异构的网络环境，采用 YANG 模型对异构网络配置协议和状态数据进行建模，形成网络领域知识图谱，用于优化网络配置管理和运行维护. 郑苏苏等人[25]考虑异质信息网络中大量网络信息缺失的问题，提出融合不完整多视图的异质信息网络嵌入方法，使得异质信息图的嵌入性能得到有效提高. 陈璐等人[26]以联邦学习作为数据治理背景，提出基于图的跨源数据错误检测模型 GEDM，为隐私保护下的数据质量维护做出贡献. 马健伟等人[27]针对海量警务办案信息的信息抽取难题，提出一种基于深度语义分析的卷宗知识抽取方法，利用 KG 和深度学习相关技术提升知识抽取性能.

上述方法[23-27]均为数字化时代的数据治理带来行之有效的深度学习解决方案. 然而，这些模型方法往往需要经历复杂的数据处理、特定的流程设计，同时需要大量的标注数据或提供足够多的专家知识以保证神经网络的拟合程度[28]. 因此，仅仅使用 KG 和小模型无法满足日益复杂的数据治理需求.

1.2 大语言模型用于数据治理

面对日渐复杂的大数据治理需求，LLM 因其强大的建模能力和良好的泛化能力，为数据治理技术带来了重大革新契机. 凭借其先进的自然语言处理能力，LLM 允许系统从非结构化数据源（如文本文件和社交媒体帖子）中提取有价值的见解.

然而，LLM 仍然存在 3 大限制：幻觉、高成本和对复杂任务执行的低准确性. 这些缺点构成了重大挑战甚至风险，阻碍了 LLM 在关键数据管理场景（如财务预测中的价格趋势分析）中的应用. 目前，已有工作[29-31]基于 LLM 智能体来克服这些挑战，如思维链和 Toolformer[32]. 这些研究工作虽然取得了一定成效，但也揭示了一些局限性. 首先，这些研究严重依赖 LLM 来支持几乎所有的任务[30]，导致不稳定性和高错误率. 其次，对于工具调用等复杂任务，需要针对特定 API 的大量训练数据来微调 LLM，带来过高的运算成本. 最后，LLM 智能体仍然缺乏充分利用来自多个来源的知识的能力. 因此，需要一种新的 LLM 增强的数据治理范式来处理这些挑战.

LLM 可以作为信息检索工具，让系统提供相关信息，而无需担心用户如何检索这些信息的复杂性. LLM 同时也是一种蕴含丰富知识的工具，已经逐步被应用于数据/知识集成[33-34]、知识库[35]、Web 数据提取[36-38]、语义 Web，甚至数据库参数调优[39]等方向. 同时，LLM 在训练和推理时间的优化方面也有大量且不断增长的工作[40-42]. 此外，LLM 大多是黑盒模型，往往不能捕获和访问事实知识. KG 作为结构化的知识模型，存储了丰富的事实知识. 但是 KG 本身难以构建和进化，这对现有的 KG 生成新事实和表示未知知识的方法提出了挑战. 如图 1 所示，我们考虑到将 LLM 和 KG 联合使用，发挥各自的优势. 一方面 KG 可以为推理提供外部知识来增强 LLM 回答的准确性，同时缓解 LLM 有关幻觉和可解释性方面的问题；另一方面，利用 LLM 对知识库中文本语料进行补全、填充

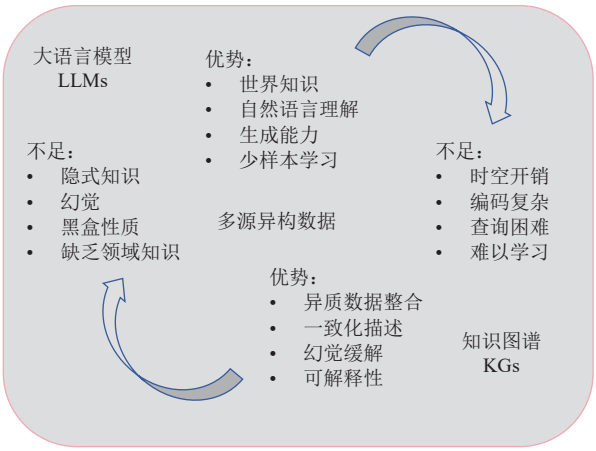


Fig. 1 Collaboration of LLMs and KGs

图 1 LLMs 和 KGs 协同

和一致性检验,解决现有 KG 方法在处理不完整的 KG 和构建 KG 时的不足.

1.3 大语言模型和 KG 协同治理

最近,LLM 和 KG 协同处理成为了数据治理领域的研究热点. LLM 大多是黑盒模型,往往不能捕获和访问事实知识. KG 作为结构化的知识模型,存储了丰富的事实知识. 但是 KG 本身难以构建和进化,这对现有的 KG 生成新事实和表示未知知识的方法提出了挑战.

LLM 和 KG 协同可以发挥各自的优势,达成互补. 一方面 KG 可以为推理提供外部知识来增强 LLM 回答的准确性,同时缓解 LLM 有关幻觉和可解释性方面的问题;另一方面,利用 LLM 对知识库中的文本语料进行补全、填充和一致性检验,解决现有 KG 方法在处理不完整的 KG 和构建 KG 时的不足.

知识图谱问答(KGQA)作为 LLM 与 KG 协同应用的核心领域[43],其目标是通过利用 KG 中存储的结构化信息来解答自然语言提出的问题. KGQA 面临的主要挑战在于如何有效地检索相关知识事实,并将 KG 的推理能力应用于问答系统[44]. 为了应对这一挑战,最新的研究趋势是利用 LLM 来桥接自然语言问题与结构化 KG 之间的差异,其中 LLM 主要扮演 2 个角色:实体和关系提取器[45-46],以及答案推理器[47-48].

尽管当前的研究集中于探索 LLM 和 KG 结合所带来的性能提升,但对于它们在处理时间和计算资源(如 token 消耗)方面的效率问题却关注不足. 因此,如何在 LLM 和 KG 的协同处理框架中优化处理性能,实现更快速的问答响应和降低成本,仍然是该领域亟待解决的研究问题. 此外,多模态 KG 能够存储现实世界中的丰富多模态信息,如何更有效地将这些异构数据集成到 LLM 中,以提高多模态知识问答的准确性,也是一个值得深入研究的领域,具有显著的发展潜力.

2 问题描述

2.1 跨域异质数据整合

对不同领域的异质数据进行高效融合是实现跨域异质数据治理的第 1 步. 过去的研究通过表示学习等方法将异质数据直接转化为词向量,并进行语义层面的对齐. 这种方法需要一定的训练成本,以及模型推理所需要的算力开销. 在本文所提出的治理框架中,跨域异质数据的语义级理解和对齐是交由 LLM 完成的. 由于 LLM 学习了大量世界知识,其对异构数据的表征能力要超越大多数传统的词嵌入模型.

在真实场景中,企业信息化往往需要多场景、多领域的异质数据,这些数据往往依托于不同的软件工具. 因此,应用数据往往源自不同的厂商,依赖于不同结构的数据库,导致相同的数据存储于异质文件中,造成信息冗余、成本高昂、拓展性低、复用性差等问题,最终导致数据孤岛现象[49],影响数据的跨领域和跨生命周期协作.

此外,跨域异质数据复杂多样的数据结构特征会影响 LLM 对数据的理解. 为提高跨域异质数据质量,规范结构特征,本文针对跨域数据的异质性和冗余性,设计 OSLC 适配器来进行跨域异质数据的高效融合. OSLC 适配器是用于解决软件数据生命周期集成问题的数据解析器[50],它可以负载于不同的商业软件中,支持大规模数据并行解析. 同时支持跨域异质数据资源的统一规范化,使其满足“链接数据”[51],支持 KG 的快速高效构建. 此外,适配器本身具有即插即用的特点,可以灵活针对不同类型的文件格式进行解析,为跨域异质数据生成统一的元数据标识. 具体地,本文给出以下定义形式:

定义 1. 跨域数据统一标识. 给定在一组异构源上存在的数据 $D=\{d,name,c,meta\}$,其中 d 表示数据所在域,c 表示数据文件的内容,$name$ 表示文件/属性名,$meta$ 表示文件元数据. 通过跨域异质数据融合算法,可以得到规范化的数据 $\bar{D}=\{id,d,name,jsc,meta,cols_index\}$. 其中 id 表示规范化的唯一标识符,d 表示该数据文件所在域,$name$ 表示数据文件名称,$meta$ 表示文件元数据,jsc 表示使用 JSON-LD 存储的文件内容. 若所存储数据为结构化数据或其他一些可以使用列式存储模型的数据格式,还会额外存储其所有属性的列索引 $cols_index$,便于快速的数据查询.

定义 2. 线性知识图[52]. 给定一个知识图谱 $\mathcal{G}$ 和一个转换后的线性知识图 $\bar{\mathcal{G}}$(knowledge line graph, KLG),线性知识图满足 2 个特征:

1) $\bar{\mathcal{G}}$ 中的 1 个节点表示 $\mathcal{G}$ 中的 1 个三元组.

2) $\bar{\mathcal{G}}$ 中的任意 2 个节点出现 1 条关联边当且仅当这 2 个节点所表示的三元组中存在公共节点.

线性知识图的存在可以很大程度上提高数据查询的效率,加速后续的数据查询算法.

2.2 跨域同源数据置信度计算

定义 3. 跨域同源数据. 对于 $\mathcal{G}$ 中的任意 2 个节点 $v_1=\{d_1,name_1,c_1,meta_1\}$ 和 $v_2=\{d_2,name_2,\ c_2,meta_2\}$,定义它们跨域同源当且仅当 $d_1\neq d_2$,$name_1=name_2,meta_1.shape=meta_2.shape$.

定义 4. 同源节点和同源子图. 对于知识图谱$\mathcal{G}$中的一组跨域同源数据$SV=\{v_i\}_{i=1}^{n}$，定义同源节点$sv=\{name,meta,num,Pr(D)\}$，同源节点集$U_{S\mathcal{G}}$，同源边集$E_{S\mathcal{G}}$. 其中$name$表示共同属性名，$meta$表示相同的文件元数据，$num$表示同源数据个数，$Pr(D)$表示数据源置信度. 定义$sv$与$v_i$的关联边$e_i=\{\omega_i\}_{i=1}^{n}$，$\omega_i$表示节点$v_i$在数据置信度计算中所占权重，则同源节点和同源知识图谱$S\mathcal{G}$构成同源子图$subS\mathcal{G}=\{U_{S\mathcal{G}}+sv,E_{S\mathcal{G}}\}$.

定义 5. 同源线性知识图. 对于知识图谱$\mathcal{G}$中的所有同源子图，构成同源知识图谱$S\mathcal{G}$. 将同源知识图谱进行线性图转换，得到同源线性知识图$S\bar{\mathcal{G}}$.

通过构造同源线性知识图，本文将跨域同源数据聚合到同一个子图中，以同源节点为核心，实现同源数据的快速一致性检验与冲突反馈. 此外，知识图谱中也存在大量孤立节点(即不存在同源数据)，这些节点也会被加入到同源线性知识图中.

2.3 多域数据查询

定义 6. 需求匹配的多域数据查询. 对于知识图谱$\mathcal{G}$的一次查询$Q(q,\mathcal{G})$，使用 LLM 对q进行意图抽取，获得查询节点(集)$v_q(V_q)$以及对应的查询信息m_q，通过同源线性知识图$S\bar{\mathcal{G}}$匹配v_q对应的同源线性子图$subS\bar{\mathcal{G}}_q$，构造线性子图的图提示嵌入 LLM，生成查询答案.

定义 7. 潜在查询冲突. 对于知识图谱$\mathcal{G}$的一次查询$Q(q,\mathcal{G})$，当q中包含的查询节点(集)$v_q(V_q)$相关信息与$\mathcal{G}$中的对应节点v的相关信息不完全一致时，出现潜在查询冲突.

3 LLM 和 KG 协同的跨域异质数据查询框架

3.1 方法框架

根据第 2 节中所阐述的 3 个问题，基于 LLM 和 KG 的跨域异质数据处理流程分为 3 个阶段. 如图 2 所示，第 1 步采用适配器进行跨域异质数据的融合和规范化，并构造相应的 KG；第 2 步使用 HKGE 算法进行线性知识子图的抽取和转化，从而对 KG 进行重构，使其更易用于检索查询；第 3 步，将用户输入 LLM 的查询内容转化为查询子图，并使用 TrustHKGM 算法进行查询节点在 KG 中的潜在冲突分析，剔除低质量节点以提高回答的可信度，并将抽取到的可信候选子图返回 LLM 中形成最终回答. 最后，本文还将上述步骤进行整合，形成多域线性知识图提示算法 MKLGP.

3.2 基于适配器的跨域异质融合算法

本文采用适配器结构进行跨域异质数据的融合，统一异质数据的存储格式. 针对实际应用场景，直接从非同源的多种数据格式中获取数据，并转化为统一的规范化表示.

遵循 OSLC 规范，本文设计 OSLC 适配器进行异质数据的解析和规范化表示. 对于给定的数据D，收集其文件名$name$、文件元数据$meta$，划分文件所在域d，同时将数据内容进行解析，使用 JSON-LD 格式进

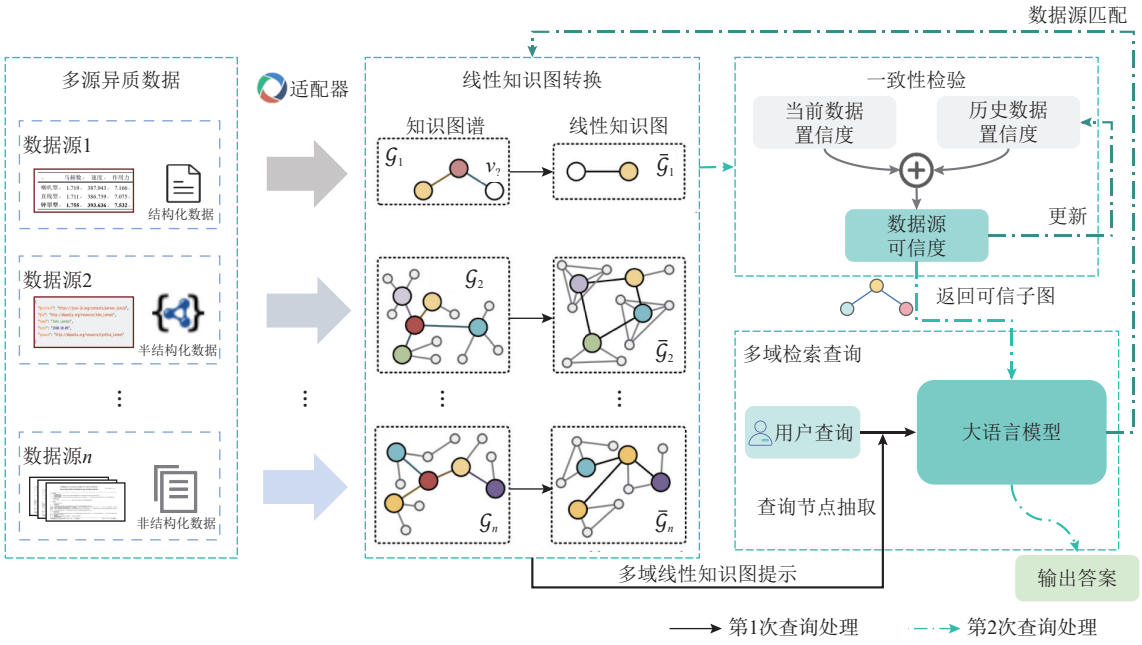


Fig. 2 A synergetic LLM-KG framework for cross-domain heterogeneous data query

图 2 LLM 和 KG 协同的跨域异质数据查询框架

行存储, 转化为链接数据. 最后赋予其唯一标识符id, 形成规范化的数据$\bar{D}$. 图 3 给出了 JSON-LD 格式的 1 个示例.

```
{
  "@context": "https://json-ld.org/contexts/person.jsonld",
  "@id": "http://dbpedia.org/resource/John_lennon",
  "name": "John Lennon",
  "born": "1940-10-09",
  "spouse": "http://dbpedia.org/resource/Cynthia_Lennon"
}
```

Fig. 3 Data format of JSON-LD

图 3 JSON-LD 数据格式

具体来说, 本文为每一种不同的数据格式都设计唯一的 OSLC 适配器来进行数据解析. 适配器的实现框架大致相同, 但需要区分结构化数据、半结构化数据和非结构化数据的解析. 对于结构化数据, 其解析相当于使用 JSON 格式存储表格数据, 文件中的属性变量采用列式存储模型(decomposition storage model, DSM)进行存储, 以便利用列索引提取所有属性信息以进行一致性检验. 对于半结构化数据, 其解析相当于使用 JSON 格式存储树形数据, 文件中采用多层嵌套 JSON 格式进行存储. 该格式数据不存在列式索引和快速查询, 需要采用树/图检索算法进行检索. 最后, 对于非结构化数据, 本文目前仅考虑文本数据, 直接进行内容存储, 并在后续考虑采用 LLM 进行实体/关系抽取任务来获得相应的信息.

最终, 跨域异质数据的融合可以表达为:

$$\begin{aligned} FusionData = &\sum_{\delta\in D_{\text{stru}}} Ada_{\text{stru}}(\delta) + \\ &\sum_{\delta\in D_{\text{semi-s}}} Ada_{\text{semi-s}}(\delta) + \\ &\sum_{\delta\in D_{\text{unstru}}} Ada_{\text{unstru}}(\delta), \end{aligned} \tag{1}$$

其中Ada_{stru}, $Ada_{\text{semi-s}}$, Ada_{unstru}分别表示结构化数据、半结构化数据和非结构化数据的适配器. D_{stru}, $D_{\text{semi-s}}$, D_{unstru}分别表示结构化数据、半结构化数据和非结构化数据的数据集合.

3.3 基于 LLM 的线性 KG 构建

完成 KG 的初步构建后, 需要遍历所有图, 获得所有的跨域同源数据组集合SV, 以及所有的孤立点集$\mathcal{L}$.

1)同源知识图抽取. 具体来说, 初始化未访问节点集合$unvisited = \{V\}$, 同源数据组集合$SV = \varnothing$, 孤立点集$\mathcal{L} = \varnothing$. 开始遍历所有节点, 依此查询各个域中的节点信息, 若匹配到同源数据, 构造同源节点sv以及对应关联e_i, 分别加入到同源节点集$U_{S\mathcal{G}}$和边集$E_{S\mathcal{G}}$, 若遍历完 1 轮后存在同源数据, 将$(U_{S\mathcal{G}}, E_{S\mathcal{G}})$加入到$SV$中; 仍未获得同源数据, 则加入孤立点集$\mathcal{L}$.

遍历完 1 轮后, 将该节点从集合$unvisited$中移除. 容易判断, 同源子图匹配的时间复杂度为$O(n\log n)$, 其中n为知识图谱$\mathcal{G}$中的节点数量. 具体的实现算法称作同源子图匹配算法 HKGE(homogeneous knowledge graph extraction), 实现如算法 1.

算法 1. 同源知识图构建算法 HKGE.

输入: 知识图谱$\mathcal{G}$, 节点集V, 边集E;

输出: 同源数据组 SV, 孤立点集$\mathcal{L}$.

① 初始化未访问节点集合$unvisited = V$, 同源数据组$SV = \varnothing$, 孤立点集$\mathcal{L} = \varnothing$;
② for each v_i in $unvisited$
③ 初始化同源节点集$U_{S\mathcal{G}} = \varnothing$, 边集$E_{S\mathcal{G}} = \varnothing$, 同源节点$sv = v_i$;
④ for each v_j in $unvisited/\{v_i\}$
⑤ if $v_j \equiv sv$
⑥ $v_j \to U_{S\mathcal{G}}, (sv, v_j) \to e_i$; /*存在同源节点, 将它们分别加入同源节点集和e_i*/
⑦ $unvisited - v_j$; /*从未访问集去除*/
⑧ end if
⑨ end for
⑩ if $U_{S\mathcal{G}} = \varnothing$
⑪ $v_i \to \mathcal{L}$; /*加入孤立点集*/
⑫ else
⑬ $e_i \to E_{S\mathcal{G}}$, $(U_{S\mathcal{G}} + sv, E_{S\mathcal{G}}) \to SV$; /*加入同源数据组*/
⑭ end if
⑮ end for

2)线性知识图的转化. 对于同源数据组集合SV中的每一个同源子图, 首先, 利用同源节点集$U_{S\mathcal{G}}$和同源边集$E_{S\mathcal{G}}$构造同源线性知识子图$subS\bar{\mathcal{G}}_i$. 将所有的$subS\bar{\mathcal{G}}_i$和孤立点集$\mathcal{L}$进行聚合, 得到同源线性知识图$S\bar{\mathcal{G}}$. 需要注意的是, $S\bar{\mathcal{G}}$仅用于同源数据的一致性检验和数据查询, 对于其他类型的查询仍然在$\mathcal{G}$中进行.

这里, 我们给出 1 个同源线性知识图的简单样例. 如图 4 所示, 同源节点与 4 个同源数据相关联, 转化为线性知识图后, 形成一个 4 阶完全图, 表明 4 个三元组之间两两同源.

3.4 跨域同源数据的置信度计算

在工程应用中, 属性变量往往在需求域被提出, 经过数据采集、仿真、实验等, 流转于多个域中. 本

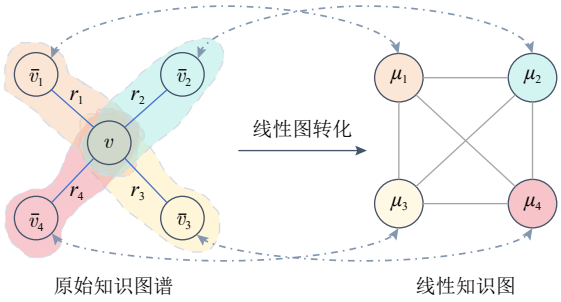


Fig. 4 Example of knowledge line graph transformation

图 4 线性知识图转化示例

文将类似这样存在于多个域中的相同来源数据定义为跨域同源数据. 这些数据需要在数据库中保证存储的一致性, 避免错误和崩溃.

本文将数据置信度的概率计算建模为最大似然估计(maximum likelihood estimation, MLE), 通过鉴别查询结果的真伪来反映数据节点会提供真值的概率. 对于实体v, 其置信度概率的计算公式为

$$\lg Pr(v)=\sum_{D\in S\hat{\mathcal{G}}}Pr(D|v)\lg\frac{Pr(v,D)}{Pr(D|v)}=\sum_{D\in S\hat{\mathcal{G}}}Pr(D|v)\lg\frac{Pr(v|D)Pr(D)}{Pr(D|v)},\tag{2}$$

其中$Pr(D|v)$和$Pr(v|D)$为 2 个条件概率. $Pr(v|D)$表示数据源 D 可靠的情况下数据节点v的数据置信度, 可以表示为:

$$Pr(v|D)=\begin{cases}Pr(D), & v\in D,\\ 1-Pr(D), & v\notin D.\end{cases}\tag{3}$$

然而, 在多域数据查询设置下, 计算所有数据源的真正例(true positive, *TP*)和假正例(false positive, *FP*)会造成大量的存算开销, 这是不合理的. 这里, 受到先前工作[53]的启发, 期望利用历史数据源的可信度和当前查询相关的数据进行$Pr(D|v)$的增量估计.

假设数据源D为所有历史查询提供的实体数为$\mathcal{H}$, 并且从源D的线性知识子图中获取的查询相关数据用$Data(q,subS\tilde{\mathcal{G}}_i)$表示. 事先假设查询实体$v_q\in Data(q,subS\tilde{\mathcal{G}}_i)$是查询$q$的正确答案之一, 那么$Data(q,subS\tilde{\mathcal{G}}_i)$中准确性分数不低于$v_q$的其他实体也可以被认为是正确的. 我们将正确答案集表示为$D_v[q]=\{v_p\in Data(q,subS\tilde{\mathcal{G}}_i)|Pr(v_p)\geqslant Pr(v_q)\}$. 然后, $Pr(D|v)$可以估计为

$$Pr(D|v)=\frac{\mathcal{H}\times Pr^h(D)+\sum_{v_p\in D_v[q]}Pr(v_p)}{\mathcal{H}+\left|Data\left(q,subS\tilde{\mathcal{G}}_i\right)\right|},\tag{4}$$

其中$Pr^h(D)$代表对数据源可信度的最新历史估计. 由于源可信度表示数据源提供正确答案的概率, 因此$\mathcal{H}\times Pr^h(D)$可以表示源$D$为历史查询提供的正确答案的预期数量. 类似地, 由于$Pr(v)$表示 1 个实体是正确答案的概率, 因此$\sum_{v_p\in D_v[q]}Pr(v_p)$表示在当前查询中由源$D$提供的正确答案数量的无偏估计. 从而, $\mathcal{H}\times Pr^h(D)+\sum_{v_p\in D_v[q]}Pr(v_p)$估计了源$D$中真正例($TP$)的数量, 分母$\mathcal{H}+|Data(q,subS\tilde{\mathcal{G}}_i)|$在一定程度上代表了源$D$提供的正样本实体的数量, 即 *TP*+*FP*.

为了确保数据置信度在[0,1]范围内, 需要规范化$Pr(v)$. 然而, 诸如 Softmax 等归一化方法会消除不同实体之间的$Pr(v)$差异, 这可能会对数据融合的性能产生负面影响. 因此, 受之前研究[54]的启发, 将$Pr(v)$转换为$-\lg(1-Pr(v))$进行归一化. 转换后, $Pr(v)$落在$(0,+\infty)$, 其中$Pr(v)$值越大表示准确性越高. 还利用 Gumbel-Softmax[55-56]来规范化$Pr(v)$, 这可以尽可能多地保留数据间的差异性. 此外, 如果大多数源为查询提供正确答案, 这意味着大多数节点均为可信节点, 则答案将更可靠. 因此, 我们引入权重系数ω_i用作投票计数, 来调整数据的准确性. 权重表示提供实体v的数据源数量N, 通常被初始化为$\frac{1}{N}$. 综上所述, 对数据置信度进行了规范化表示为:

$$Pr(\tilde{v})\leftarrow\frac{\exp\left(\frac{-\omega_v\times\lg(1-Pr(v))}{z}\right)}{\sum_{v\in Data\left(q,subS\tilde{\mathcal{G}}_i\right)}\exp\left(\frac{-\omega_v\times\lg(1-Pr(v))}{z}\right)},\tag{5}$$

其中z是一个温度控制参数. 当z接近 0 时, softmax 计算相当于求极大值; 而当z接近$+\infty$时, 归一化数据准确性的分布与均匀分布相同.

为了估计$Pr(D|v)$, 采用 *TP* 和 *FP* 重新定义数据源可信度. 对于每个数据源发出正确查询答案的真预测或假预测, 当 LLM 生成的实体本身置信度较高, 即$Pr(v)$较高时, 预测可以被视为真预测. 据此, 数据源可信度是指从数据源中发出的正样本实体是真正正确(TP)的概率, 它由$\frac{TP}{TP+FP}$估计, 其中 *TP* 表示提供的正确答案的数量, *FP* 表示提供的错误答案的数量.

定义数据置信度来说明同源知识子图中的每个节点对会提供真值的概率. 由于$Pr(v)$和$Pr(D|v)$均在式(2)(4)中给出, 根据全概率公式, 对$Pr(D)$的估计可以扩展为

$$Pr(D)=\sum_{v\in Data\left(q,subS\hat{\mathcal{G}}_i\right)}Pr(D|v)\times Pr(\tilde{v}),\tag{6}$$

其中条件概率$Pr(D|v)$表示当实体v是查询的真实答案时数据源D的可信度得分. $Pr(\tilde{v})$表示实体v经过规

范化后的数据准确性. $Data(q, subS\bar{\mathcal{G}}_i)$表示从数据源$D$对应的线性知识子图中获取的查询相关数据集合.

最终, 我们设计可信候选子图匹配算法TrustHKGM, 对同源子图的数据源可信度进行计算, 以保证嵌入LLM的知识图质量. 该算法首先初始化每个节点的置信度$Pr(\upsilon)$, 并计算$Pr(D|\upsilon)$, 根据式(6)得到$Pr(D)$; 接下来遍历子图中的每个节点, 更新其预估的节点置信度, 并剔除不可信节点. 算法2的实现流程如下.

算法2. 可信候选子图匹配算法TrustHKGM.

输入: 数据源D, 同源线性知识图$S\bar{\mathcal{G}}$;

输出: 数据源可信度$Trust_D$, 可信候选子图$T\text{-}subS\bar{\mathcal{G}}_i$.

① 初始化$S\bar{\mathcal{G}}$中所有数据的置信度$Trust_0$, 数据源可信度的最新历史估计$Pr^h(D)$为$Pr(D)$, 可信候选子图 $T\text{-}subS\bar{\mathcal{G}}_i$;
② for each $subS\bar{\mathcal{G}}_i$ in $S\bar{\mathcal{G}}$
③ 初始化 $N\rightarrow|subS\bar{\mathcal{G}}_i|$;
④ 根据式(3)计算$Pr(\upsilon|D)$; /*如果是与数据源匹配的子图, 则其值为$Pr(D)$, 否则为$1-Pr(D)$*/
⑤ 根据式(4)计算更新$Pr(D|\upsilon)$; /*查询实体υ_q的置信度是$Pr(D)$的值浮动0.000 5, 由LLM决定*/
⑥ 根据式(6)计算更新$Pr(D)$;
⑦ for each υ in $subS\bar{\mathcal{G}}_i$
⑧ 根据式(2)计算更新$Pr(\upsilon)$;
⑨ 根据式(5)计算更新$Pr(\tilde{\upsilon})$;
⑩ if $Pr(\tilde{\upsilon})<Pr(D)$
⑪ $subS\bar{\mathcal{G}}_i-\upsilon$; /* 当节点可信度低于源可信度, 认为其不可信, 从子图中删除*/
⑫ end if
⑬ end for
⑭ $subS\bar{\mathcal{G}}_i\rightarrow T\text{-}subS\bar{\mathcal{G}}_i$;
⑮ $Pr(\tilde{\upsilon})\rightarrow Trust_D$;
⑯ end for

3.5 线性知识图提示设计

本文提出用于多域异构数据查询的多域线性知识图提示(multi-domain knowledge line graph prompting, MKLGP)算法. 给定一个用户的查询q, 我们首先使用LLM对q进行意图抽取, 获得查询节点集$V_q(\upsilon_q, m_q)$, 其中υ_q表示查询的可能节点, m_q表示对应的查询信息.

用户的查询可能只包含υ_q和m_q中的1个, 此时仅通过该信息在同源线性知识图$S\bar{\mathcal{G}}$中进行查找匹配, 并返回所有符合条件的线性同源子图$subS\bar{\mathcal{G}}_i$和孤立点集$\mathcal{L}$所构成的答案子图$\mathcal{G}_A$. 若用户的查询同时包含υ_q和m_q, 则分别进行2次查询匹配, 得到2个答案候选子图$\mathcal{G}_{A_q}$和$\mathcal{G}_{A_m}$, 并取交集得到最终答案子图$\mathcal{G}_A$.

算法3给出了MKLGP算法的具体实现. 用户查询首先输入到LLM进行解析抽取, 得到V_q后, 构造图提示模板进行V_q的查询和匹配, 返回未经检验的答案子图$\mathcal{G}_A$. 之后对$\mathcal{G}_A$进行一致性检验, 使用TrustHKGM算法剔除其中的不可信/冲突节点, 并返回可信答案子图$T\text{-}sub\bar{\mathcal{G}}_s$, 并作为LLM的输入, 用于输出最终的查询结果.

算法3. 多域线性知识图提示算法MKLGP.

输入: 用户查询q;

输出: 查询答案Ans.

① 构造少样本提示, 对查询q执行意图识别和实体抽取任务, 得到结构化查询集合Q_0;
② 初始化查询节点集$V_q\leftarrow Q_0$;
③ 初始化可信答案集$Ans_T\leftarrow\varnothing$;
④ 根据算法1获得同源线性知识图$S\bar{\mathcal{G}}$;
⑤ for each D in V_q
⑥ 根据算法2获得数据源可信度 $Trust_D$;
⑦ 根据算法2获得可信候选子图 $T\text{-}sub\bar{\mathcal{G}}_s$;
⑧ $Ans_T\leftarrow(Trust_D, T\text{-}sub\bar{\mathcal{G}}_s)$; /* 将2项数据存入可信答案集合, 后续反馈给LLM */
⑨ end for
⑩ 根据q和Ans_T构造线性知识图提示, 输入LLM进行2次提问;
⑪ 输出Ans.

4 实 验

4.1 实验设置

1)数据集. 实验是在4个真实世界的基准数据集上进行的[54,57-58]. 表1展示了这些数据集的具体信息. ①电影数据集, 包含从13个来源收集的电影数据. 本文对其发出了210个查询. ②图书数据集, 包含来自10个来源的图书数据. 本文对其发出了100个查询. ③航班数据集, 从20个来源收集了1 200多个航班的信息. 本文对其发出了260个查询. ④股票数据集, 收集来自20个来源的1 000个股票符号的交易数据. 本文对其发出了100个查询.

2)实现细节. 所有的方法都是在Python 3.8中实现, 查询使用的大语言模型基座为Llana3-8B-Instruct-Chat. 每一类数据格式都配备对应适配器来转化为

Table 1 Statistics of the Preprocessed Datasets

表 1 预处理后的数据集信息统计

数据集	数据源（缩写）	源数量	实体数量	关系数量	查询数
电影	JSON(J)	4	19 701	45 790	210
	KG(K)	5	100 229	264 709	
	CSV(C)	4	70 276	184 657	
书籍	JSON(J)	3	3 392	2 824	100
	CSV(C)	3	2 547	1 812	
	XML(X)	4	2 054	1 509	
航班	CSV(C)	10	48 672	100 835	260
	JSON(J)	10	41 939	89 339	
股票	CSV(C)	10	7 799	11 169	100
	JSON(J)	10	7 759	10 619	

JSON-LD 数据格式，并以此来存储对应的知识图谱信息. 对于超参数设置，温度参数z设置为 0.5，学习速率θ设置为 3E−5. 历史查询的实体数量$\mathcal{H}$初始化为 50，初始的数据置信度和数据源置信度$Pr(D)$均定义为 0.8. 所有的实验都是在 1 台拥有 Intel® Core™ Ultra 9 185H 2.30 GHz 和内存 512 GB 的设备上进行的.

此外，由于v_q的作用是进行$Data(q, subS\bar{\mathcal{G}}_i)$中实体的粗分类任务，本文利用 GPT-3.5 Turbo 进行数据源参照的查询实体v_q的置信度值的生成，定义其值在未更新的$Pr(D)$值浮动 0.000 5，以保证查询过程中置信度变化的鲁棒性.

3）评价指标. 为了评估有效性，我们参考先前的实验指标[56-58]，使用 $F1$ 分数作为数据融合结果的评价指标，即精度（precision，P）和召回率（recall，R）的调和平均值，计算公式为 $F1=\dfrac{2P\times R}{P+R}$. 此外，我们还使用运行时间 T 作为一个评价指标来验证效率.

4.2 算法鲁棒性和效率分析

本文从 2 个方面进行实验来分别证明一致性检验的有效性和线性 KG 转化对查询性能的提升：1）对于 4 个预处理过后的数据集，进行 30%~50% 的随机的实体和关系交换，以破坏跨域数据的一致性. 之后，我们采用 TrustHKGM 算法对所有同源节点进行匹配，以观测我们的方法在不同质量数据集上的鲁棒性. 2）在相同的实验数据中，对常规 KG 和转化后的线性知识图谱（KLG）均采用 TrustHKGM，对比转化前后的数据处理在时间和性能上的差异性.

LLM 的 token 开销主要集中于查询实体（查询图）的抽取和生成、查询实体置信度的评估以及最终的结果生成. 实验结果表明，不同数据集的 token 开销如下：电影数据集约为 84 000 token，书籍数据集约为 40 000 token，航班数据集约为 104 000 token，股票数据集约为 40 000 token.

4.3 多域数据查询实验对比

基线模型比较. 为了证明多域数据检索算法的优越性，本文与其他 SOTA 方法进行了比较，包括多文档问答方法 MD-QA、数据融合查询方法 FusionQuery[53]、基于知识库的问答方法 ChatKBQA. 同时，还与一些基线方法进行比较，包含基于迭代的数据查询方法 TruthFinder[57]、概率数据检索方法 LTM[58].

采用上述的 SOTA 方法和基线模型来替换本文提出的查询算法，可以获得 5 组实验结果来评估其多域查询的性能. 表 2 展示了各类方法在 4 个数据集上的数据查询性能，即 $F1$ 得分和查询时间.

Table 2 Comparison of Baseline Models and SOTA Models by Effectiveness and Efficiency

表 2 从性能和效率角度与基线模型和 SOTA 模型的对比实验

数据集	数据类型	基础模型				SOTA 模型						本文方法	
		LTM		TruthFinder		ChatKBQA		MD-QA		FusionQuery		MKLGP	
		$F1$/%	时间/s	$F1$/%	时间/s	$F1$/%	时间/s	$F1$/%	时间/s	$F1$/%	时间/s	$F1$/%	时间/s
电影	J, K	41.4	1 995	37.1	9 717	43.2	3 809	46.2	1 588	**53.2**	122.4	52.6	**98.3**
	J, C	42.9	1 884	41.9	7 214	45.0	3 246	44.5	1 360	52.7	183.1	**54.3**	**75.1**
	K, C	41.2	1 576	37.8	2 199	37.6	2 027	45.2	987	42.5	141.0	**49.1**	**86.0**
	J, K, C	40.8	2 346	36.6	11 225	41.5	5 151	49.8	2 264	53.6	137.8	**54.8**	**157**
书籍	J, C	42.4	195.3	40.2	1 017	35.2	165.0	55.7	14.2	58.5	22.7	**62.5**	**3.66**
	J, X	35.6	277.7	35.5	1 070	36.1	200.1	55.1	15.6	57.9	20.6	**61.1**	**3.78**
	C, X	44.1	232.6	43.0	1 033	42.6	201.4	57.2	15.6	**60.3**	21.5	59.0	**3.54**
	J, C, X	41.0	413.2	37.3	2 304	40.4	394.1	56.4	22.6	59.1	27.0	**59.8**	**7.4**
航班	C, J	79.1	14 786	27.3	6 049	72.3	376	**76.5**	160	74.2	**20.2**	72.9	180
股票	C, J	19.2	1 337	68.4	2.30	64.8	88.9	65.2	78.4	68.0	**0.33**	**74.6**	12.1

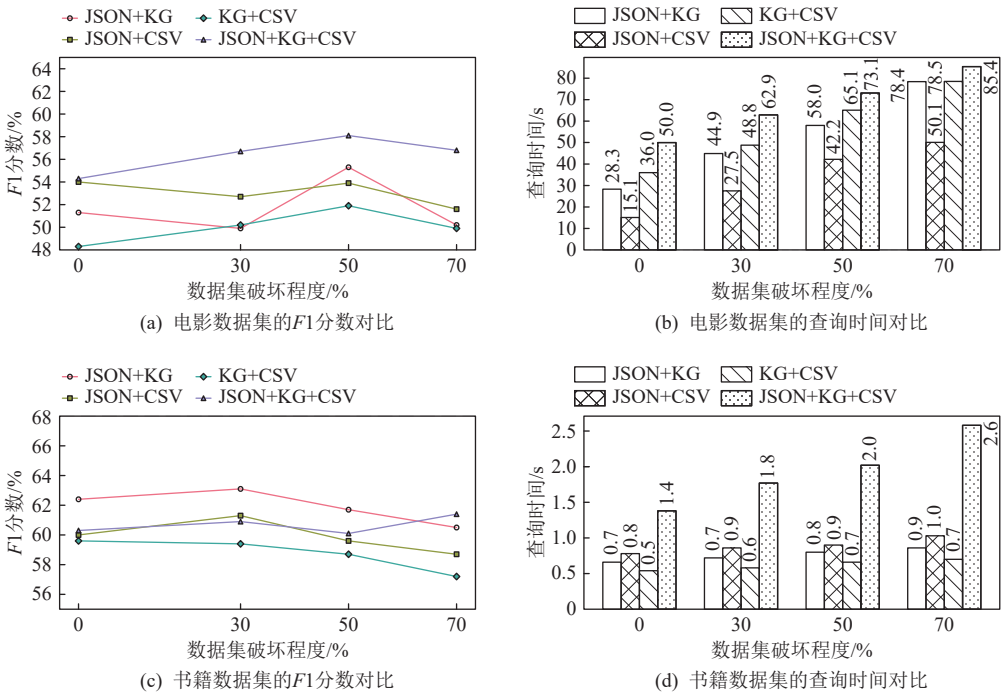
注：黑体数值表示最优指标.

4.4 结果分析

4.4.1 一致性检验分析

直观上说, KLG 将跨域同源数据进行了聚类, 仅需检索少量中心节点就可直接获得所有同源数据的信息, 并展开检测算法. 因此大幅减少了在传统知识图谱中遍历游走查询带来的时间损耗.

对于一致性检验算法的性能实验, 本文对数据集分别采取了 30%, 50%, 70% 的数据集破坏, 并使用 TrustHKGM 算法进行数据冲突检测, 破坏规则分为 MASK 遮盖和随机替换. 图 5 展示了未破坏数据集和 3 个破坏程度下的查询时间变化和准确率变化. 观察图 5 可以发现, 随着数据集质量的降低, 整个检索流程的时间会变得更长. 这是因为子图匹配过程中冲突子图变多, 生成了更多的图嵌入文本, 导致 LLM 需要更多时间来进行推理和问答. 在准确性层面, 检索的 *F*1 分数则保持一定程度稳定. 虽然也有少量精度降低, 但整体上, 因为筛选了大量低质量的图谱节点, 依然能保证问答的准确性.



(a) 电影数据集的*F*1分数对比 (b) 电影数据集的查询时间对比

(c) 书籍数据集的*F*1分数对比 (d) 书籍数据集的查询时间对比

Fig. 5 Perfomance comparison under scenarios of 0%, 30%, 50% and 70% corruption in the movie and book datasets

图 5 在电影和书籍数据集中发生 0%, 30%, 50%, 70% 的破坏场景下的性能比较

此外, 针对 TrustHKGM 算法展开消融实验. 观察图 5(c)折线图可以发现, 当不采用该算法进行子图筛选而直接检索时, 检索的 *F*1 分数始终较低, 这是因为数据集中本身存在一定量的噪声数据, 影响检索的性能. 而随着破坏程度上升, 数据噪声增大和 LLM 检索到的大部分数据均是错误的, 导致了较低的 *F*1 分数. 此外, 不采用 TrustHKGM 算法时, 提示词中嵌入的知识图谱也未经筛选, 导致检索量增大和检索时间变长.

同时, 表 3 展示了线性知识图对整个查询处理方法的性能影响. 以 *F*1 分数作为查询性能的评价指标, 以查询时间(query time, QT)和数据处理时间(processing time, PT)作为查询效率的评价指标, 并作差给出直观的收益展示.

从表 3 中可以看出, 线性知识图的引入在性能和检索效率上均能带来显著提升. 在性能方面, 几乎所有数据集的所有数据域分布上均能带来 *F*1 分数的提升. 在检索效率上, KLG 的构建确实会增加一定的数据处理时间, 对于数据规模最庞大的航班数据集, 其构建时间甚至超过了 100 s, 但其对查询时间的降低更为明显. 从小规模的书籍数据集到大规模的电影航班数据集, 其查询时间可以压缩超 100 倍. 其中航班数据集的查询时间提升被 NAN(not a number)代替, 以彰显 29.8 s 和 13.2 h 之间的庞大差距. 不难看出, KLG 的构建允许系统仅需检索少量的同源节点即可覆盖全域图谱的查询, 使其拥有落地部署的潜力.

4.4.2 多域检索查询结果分析

表 2 显示, MKLGP 算法在 4 个数据集上均优于所有的对比模型. 实验结果表明, 它比 TruthFinder 和 LTM 的 *F*1 分数高了近 1.5 倍和 2 倍, 并且比其他基

Table 3 Ablation Experiments of KLG

表 3 KLG 的消融实验

数据集	数据源	MKLGP			-KLG			差值		
		*F*1/%	QT/s	PT/s	*F*1/%	QT/s	PT/s	*F*1/%	QT/s	PT/s
电影	J, K	51.3	25.7	2.64	12.2	2 783	0.28	**−39.1**	**+2 757.3**	−2.36
	J, C	54.0	12.7	2.36	49.1	1 882	0.29	**−4.9**	**+1 869.3**	−2.07
	K, C	48.3	31.6	4.40	45.5	4 233	0.29	**−2.8**	**+4 291.4**	−4.11
	J, K, C	54.3	39.2	10.8	50.5	4 437	0.32	**−3.8**	**+4 397.8**	−10.48
书籍	J, C	62.4	0.19	0.47	57.1	11.9	0.17	**−5.3**	**+11.71**	−0.3
	J, X	60.0	0.22	0.56	59.3	11.7	0.17	**−0.7**	**+11.48**	−0.39
	C, X	59.4	0.16	0.38	55.3	8.39	0.16	**−4.1**	**+8.23**	−0.22
	J, C, X	60.3	0.31	1.07	57.2	15.8	0.18	**−3.1**	**+15.49**	−0.89
航班	C, J	72.9	29.8	109.9	75.2	13.2h	0.5	+2.1	**NAN**	−109.4
股票	C, J	71.6	0.72	0.36	69.6	450.8	0.19	**−2.0**	**+450.02**	−0.17

注：黑体数值表示最优指标. QT 表示查询时间，PT 表示数据处理时间，MKLGP 表示采用多域线性图提示嵌入算法的性能分析，-KLG 表示不使用线性知识图情况下的性能分析.

线获得了更好的性能. TruthFinder 方法在所有数据集上的性能都较差，这是因为 TruthFinder 只能为查询返回 1 个答案，无法适应 1 个查询通常有多个返回值的情况. 如，1 部电影或 1 本书通常有多个导演或作者. LTM 并非在所有数据集上都表现良好. 这是因为它们需要从大数据中学习数据的准确性，而这在低资源环境下是不可用的.

对于 SOTA 方法，选择了 FusionQuery, MD-QA 和 ChatKBQA 领域在 2024 年下的最新研究成果. 虽然这些方法并非专注于低资源、高噪声的数据场景，但它们在该场景下仍然具有较好的鲁棒性和检索性能. MD-QA 和 FusionQuery 采用 BM25 检索器和 Transformer 架构的解码器进行答案生成；而 ChatKBQA 模型采用了基于 LLM 的数据检索方法，时空开销主要集中于 token 的消耗和基于 LLM 的检索.

对比可知，MKLGP 算法将开销集中在 KLG 的构建上. 虽然在 KLG 的原文中，其构建时间往往在几秒以内，十分高效，但引入了 LLM 后仍然会因为文本生成而带来额外的时间开销，但这仍然在可接受范围内. 最后，这些方法均展现出较好的检索性能，但由于数据集本身存在大量的噪声，其问答准确率的提升将变得十分有限.

MKLGP 算法在大多数情况下都达到最好或类似的性能. 在处理单个电影、图书、航班和股票数据时，MKLGP 算法分别的平均工作流程完成时间为 1.42 s, 1.06 s, 0.64 s, 0.66 s. 这一效率优于大多数实际数据检索引擎的更新周期. 此外，即使在数据更新频率高于查询执行时间的极端情况下，该算法依然展现出了卓越的鲁棒性，有效避免了数据更新带来的潜在影响.

4.4.3 局限性分析

MKLGP 在同源知识图构建和子图匹配模块仍然存在较大的性能开销，且无法针对广泛的多模态知识库或复杂的非结构化数据进行问答，仍需额外的算法来接入到完整的 KBQA 系统中.

5 总结与展望

本文提出了基于 LLM 和 KG 的跨域异质数据处理方法，它包括跨域异质数据融合、一致性检验和多域数据检索 3 个模块. 在数据融合阶段，通过适配器整合跨域异质数据，构造相应的知识图谱，并采用线性知识图技术对图谱进行优化. 在一致性检验阶段，对查询相关数据进行筛选与匹配. 在多域数据查询阶段，所提出的 MKLGP 算法也很好地构建基于 LLM 的快速高效查询方法. 综合实验证明所提方法的优越性.

未来计划进一步提高跨域异质处理的有效性和高效性. 包括将方法框架推广到向量数据库中，嵌入到 RAG 中实现一体化执行管道，构建低成本、轻量级的多域检索查询体系. 同时，考虑数据源不均衡的非稳态场景下基于 LLM 的跨域异质数据管理方案，也是值得进一步探索的方向.

作者贡献声明：吴文隆提出了算法思路和实验方案，并完成论文撰写；尹海莲、王昊奋、李博涵、丁岩负责完成项目统筹和写作指导；王宁负责算法优化；徐梦飞负责论文实验设计和结果分析；赵鑫

喆和殷崭祚负责模型优化以及论文润色；刘元睿负责实验数据预处理和分析. 李博涵（bhli@nuaa.edu.cn）和丁岩（dingyan@cmss.chinamobile.com）为共同通信作者.

参 考 文 献

[1] National Development and Reform Commission. Interpretation of the "14th Five-Year Plan" for the development of the digital economy | accelerating the market-oriented construction of data elements and fully utilizing the role of data elements[EB/OL]. (2022-01-03)[2024-07-15]https://www.ndrc.gov.cn/xxgk/jd/jd/202201/t20220121_1312584.html
(中华人民共和国国家发展和改革委员会.《"十四五"数字经济发展规划》解读｜加快推进数据要素市场化建设 充分发挥数据要素作用[EB/OL].(2022-01-03)[2024-07-15]. https://www.ndrc.gov.cn/xxgk/jd/jd/202201/t20220121_1312584.html)

[2] Du Xiaoyong, Li Tong, Lu Wei, et al. Cross-domain data management[J]. Computer Science, 2024, 51(1): 4−12 (in Chinese)
(杜小勇, 李彤, 卢卫, 等. 跨域数据管理[J]. 计算机科学, 2024, 51(1): 4−12)

[3] Yan Jiahe, Li Honghui, Ma Ying, et al. Multi-source heterogeneous data fusion technologies and government big data governance system[J]. Computer Science, 2024, 51(2): 1−14 (in Chinese)
(闫佳和, 李红辉, 马英, 等. 多源异构数据融合关键技术与政务大数据治理体系[J]. 计算机科学, 2024, 51(2): 1−14)

[4] Qin Yuan, Ye Yuan, Zhenyu Wen, et al. An effective framework for enhancing query answering in a heterogeneous data lake[C]//Proc of the 46th Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2023: 770−780

[5] Wu Xindong, Zhu Xingquan, Wu Gongqing, et al. Data mining with big data[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(1): 97−107

[6] Wang Jun, Wang Xiulai, Pang Wei, et al. Research on big data governance for science and technology forecast[J]. Computer Science, 2021, 48(9): 36−42 (in Chinese)
(王俊, 王修来, 庞威, 等. 面向科技前瞻预测的大数据治理研究[J]. 计算机科学, 2021, 48(9): 36−42)

[7] Labrinidis A, Jagadish H V. Challenges and opportunities with big data[J]. Proceedings of the VLDB Endowment, 2012, 5(12): 2032−2033

[8] Yang Jia, Huang Fang, Long Jun, et al. Heterogeneous data conversion based on semantic models of expert information[J]. Computer Systems Applications, 2010, 19(10): 57−62 (in Chinese)
(杨佳, 黄芳, 龙军, 等. 专家信息语义模型异构数据转换技术[J]. 计算机系统应用, 2010, 19(10): 57−62)

[9] Popa L, Velegrakis Y, Miller R J, et al. Translating web data[C]//Proc of the 28th Int Conf on Very Large Databases. New York: ACM, 2002: 598−609

[10] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint, arXiv: 1301.3781, 2013

[11] Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations[C]//Proc of the 2013 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2013: 746−751

[12] Mikolov T, Sutskever I, Chen Kai, et al. Distributed representations of words and phrases and their compositionality[C]//Proc of the 27th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2013: 26

[13] Wang Meng, Wang Haofen, Li Bohan, et al. Survey on key technologies of new generation knowledge graph[J]. Journal of Computer Research and Development, 2022, 59(9): 1947−1965 (in Chinese)
(王萌, 王昊奋, 李博涵, 等. 新一代知识图谱关键技术综述[J]. 计算机研究与发展, 2022, 59(9): 1947−1965)

[14] Chen Huimin, Liu Zhiyuan, Sun Maosong. The social opportunities and challenges in the era of large language models[J]. Journal of Computer Research and Development, 2024, 61(5): 1094−1103 (in Chinese)
(陈慧敏, 刘知远, 孙茂松. 大语言模型时代的社会机遇与挑战[J]. 计算机研究与发展, 2024, 61(5): 1094−1103)

[15] Hong Sirui, Lin Yizhang, Liu Bang, et al. Data interpreter: An LLM agent for data science[J]. arXiv preprint, arXiv: 2402.18679, 2024

[16] Hu Songlin, Li Juanzi, Qin Bing, et al. The dual nature of large models: An introduction to the special topic on large models and security[J]. Journal of Computer Research and Development, 2024, 61((5):): 1085−1093 (in Chinese)
(虎嵩林, 李涓子, 秦兵, 等. 亦正亦邪大语言模型——大语言模型与安全专题导读[J]. 计算机研究与发展, 2024, 61(5): 1085−1093)

[17] Pan Shirui, Luo Linhao, Wang Yufei, et al. Unifying large language models and knowledge graphs: A roadmap[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(7): 3580−3599

[18] Feng Yangyang, Wang Qing, Shu Jiwu, et al. From BERT to ChatGPT: Challenges and technical development of storage systems for large model training[J]. Journal of Computer Research and Development, 2024, 61(4): 809−823 (in Chinese)
(冯杨洋, 汪庆, 舒继武, 等. 从 BERT 到 ChatGPT: 大语言模型训练中的存储系统挑战与技术发展[J]. 计算机研究与发展, 2024, 61(4): 809−823)

[19] Zhu Hongyin, Peng Hao, Lyu Zhiheng, et al. Pre-training language model incorporating domain-specific heterogeneous knowledge into a unified representation[J]. Expert Systems with Applications, 2023, 215(1): 119369

[20] Hu Linmei, Liu Zeyi, Zhao Ziwang, et al. A survey of knowledge enhanced pre-trained language models[J]. IEEE Transactions on Knowledge and Data Engineering, 2024, 36(4): 1413−1430

[21] Ji Ziwei, Lee N, Frieske R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1−38

[22] Zhu Di, Zhang Bowen, Cheng Yaqi, et al. Survey on knowledge enabled new generation information systems[J]. Journal of Software, 2023, 34(10): 4439−4462 (in Chinese)
(朱迪, 张博闻, 程雅琪, 等. 知识赋能的新一代信息系统研究现状、发展与挑战[J]. 软件学报, 2023, 34(10): 4439−4462)

[23] Yang Xiaohui, Wan Rui, Zhang Haibin, et al. Semantical symbol mapping embedding learning algorithm for knowledge graph[J]. Journal of Computer Research and Development, 2018, 55(8): 1773−1784 (in Chinese)
(杨晓慧, 万睿, 张海滨, 等. 基于符号语义映射的知识图谱表示学习算法[J]. 计算机研究与发展, 2018, 55(8): 1773−1784)

[24] Dong Yongqiang, Wang Xin, Liu Yongbo, et al. Building network domain knowledge graph from heterogeneous YANG models[J]. Journal of Computer Research and Development, 2020, 57(4): 699−708 (in Chinese)
(董永强, 王鑫, 刘永博, 等. 异构 YANG 模型驱动的网络领域知识图谱构建[J]. 计算机研究与发展, 2020, 57(4): 699−708)

[25] Zheng Susu, Guan Donghai, Yuan Weiwei. Heterogeneous information network embedding with incomplete multi-view fusion[J]. Computer Science, 2021, 48(9): 68−76 (in Chinese)
(郑苏苏, 关东海, 袁伟伟. 融合不完整多视图的异质信息网络嵌入方法[J]. 计算机科学, 2021, 48(9): 68−76)

[26] Chen Lu, Guo Yuxiang, Ge Congcong, el al. Cross-source data error detection approach based on federated learning[J]. Journal of Software, 2023, 34(3): 1126−1147 (in Chinese)
(陈璐, 郭宇翔, 葛丛丛. 基于联邦学习的跨源数据错误检测方法[J]. 软件学报, 2023, 34(3): 1126−1147)

[27] Ma Jianwei, Wang Tiexin, Jiang Hong, et al. Knowledge extraction based on deep semantics analysis towards police dossier[J]. Journal of Computer Research and Development, 2024, 61(5): 1325−1335 (in Chinese)
(马健伟, 王铁鑫, 江宏, 等. 基于深度语义分析的警务卷宗知识抽取[J]. 计算机研究与发展, 2024, 61(5): 1325−1335)

[28] Tu Jianhong, Fan Ju, Tang Nan, et al. Unicorn: A unified multi-tasking model for supporting matching tasks in data integration[J]. Proceedings of the ACM on Management of Data, 2023, 1(1): 1−26

[29] Shinn N, Labash B, Gopinath A. Reflexion: An autonomous agent with dynamic memory and self-reflection[J]. arXiv preprint, arXiv: 2303.11366, 2023

[30] Wei Jason, Wang Xuezhi, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models[C]// Proc of the 36th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 35: 24824−24837

[31] Yao Shunyu, Zhao Jeffrey, Yu Dian, et al. React: Synergizing reasoning and acting in language models[J]. arXiv preprint, arXiv: 2210.03629, 2022

[32] Schick T, Dwivedi-Yu J, Dessì R, et al. Toolformer: Language models can teach themselves to use tools[C]//Proc of the 38th Int Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2024: 36

[33] Doan A H, Halevy A Y. Semantic integration research in the database community: A brief survey[J]. AI Magazine, 2005, 26(1): 83−83

[34] Dong X L. Challenges and innovations in building a product knowledge graph[C]//Proc of the 24th SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 2869−2869

[35] De Sa C, Ratner A, Ré C, et al. Deepdive: Declarative knowledge base construction[J]. ACM SIGMOD Record, 2016, 45(1): 60−67

[36] Etzioni O, Cafarella M, Downey D, et al. Unsupervised named-entity extraction from the web: An experimental study[J]. Artificial Intelligence, 2005, 165(1): 91−134

[37] Madhavan J, Jeffery S R, Cohen S, et al. Web-scale data integration: You can only afford to pay as you go[C]//Proc of the 3rd Biennial Conference on Innovative Data Systems Research. New York: ACM, 2007: 342−350

[38] Cafarella M J, Madhavan J, Halevy A. Web-scale extraction of structured data[J]. ACM SIGMOD Record, 2009, 37(4): 55−61

[39] Trummer I. The case for NLP-enhanced database tuning: Towards tuning tools that “read the manual”[J]. Proceedings of the VLDB Endowment, 2021, 14(7): 1159−1165

[40] Miao Xupeng, Wang Yujie, Jiang Youhe, et al. Galvatron: Efficient transformer training over multiple gpus using automatic parallelism[J]. arXiv preprint, arXiv: 2211.13878, 2022

[41] Um T, Oh B, Seo B, et al. Fastflow: Accelerating deep learning model training with smart offloading of input data pipeline[J]. Proceedings of the VLDB Endowment, 2023, 16(5): 1086−1099

[42] Zhao Yanlin, Gu A, Varma R, et al. Pytorch fsdp: Experiences on scaling fully sharded data parallel[J]. arXiv preprint, arXiv: 2304.11277, 2023

[43] Tan Y, Min D, Li Y, et al. Can ChatGPT replace traditional KBQA models? An in-depth analysis of the question answering performance of the GPT LLM family[C]//Proc of the 22nd Int Semantic Web Conf. Berlin: Springer, 2023: 348−367

[44] Luo Haoran, E haihong, Tang Zichen, et al. ChatKBQA: A generate-then-retrieve framework for Kkowledge base question answering with fine-tuned large language models[C]//Findings of the 62nd Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 2039−2056

[45] Hu Nan, Wu Yike, Qi Guilin, et al. An empirical study of pre-trained language models in simple knowledge graph question answering[C]// Proc of the 30th Int Conf on World Wide Web. New York: ACM, 2023, 2855−2886

[46] Xu Yichong, Zhu Chenguang, Xu Ruochen, et al. Fusing context into knowledge graph for commonsense question answering[C]//Findings of the 59th Association for Computational Linguistics. Stroudsburg, PA: ACL, 2021: 1201−1207

[47] Jiang Jinhao, Zhou Kun, Zhao W X, et al. Unikgqa: Unified retrieval and reasoning for solving multi-hop question answering over knowledge graph[J]. arXiv preprint, arXiv: 2212.00959, 2022

[48] Jiang Jinhao, Zhou Kun, Dong Zican, et al. Structgpt: A general framework for large language model to reason over structured data[J]. arXiv preprint, arXiv: 2305.09645, 2023

[49] Fernandez R C, Elmore A J, Franklin M J, et al. How large language models will disrupt data management[J]. Proceedings of the VLDB Endowment, 2023, 16(11): 3302−3309

[50] Fabio B, Bruno M S F, Rafael T, et al. Model-driven integration and the OSLC standard: A mapping of applied studies[C]//Proc of the 38th ACMIGAPP Symp on Applied Computing. New York: ACM, 2023: 763−770

[51] Bizer C, Heath T, Idehen K, et al. Linked data on the web [C]//Proc of the 17th Int Conf on World Wide Web. New York: ACM, 2008: 1265−1266

[52] Fionda V, Pirrò G. Learning triple embeddings from knowledge graphs[C]//Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA : AAAI, 2020, 3874−3881

[53] Zhu Junhao, Mao Yuren, Chen Lu, et al. FusionQuery: On-demand fusion queries over multi-source heterogeneous data[J]. Proceedings of the VLDB Endowment, 2024, 17(6): 1337−1349

[54] Yin Xiaoxin, Han Jiawei, Yu P S. Truth discovery with multiple conflicting information providers on the web[C]//Proc of the 13th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2007: 1048−1052

[55] Jang E, Gu Shixiang, Poole B. Categorical reparameterization with gumbel-softmax[J]. arXiv preprint, arXiv: 1611.01144, 2016

[56] Dong X L, Berti-Equille L, Srivastava D. Integrating conflicting data: The role of source dependence[J]. Proceedings of the VLDB Endowment, 2009, 2(1): 550−561

[57] Li Xian, Dong X L, Lyons K, et al. Truth finding on the deep web: Is the problem solved?[J]. Proceedings of the VLDB Endowment, 2012, 6(2): 97−108

[58] Zhao Bo, Rubinstein B I P, Gemmell J, et al. A Bayesian approach to discovering truth from conflicting sources for data integration[J]. Proceedings of the VLDB Endowment, 2012, 5(6): 550−561

[59] Wang Yu, Lipka N, Rossi R A, et al. Knowledge graph prompting for multi-document question answering[C]//Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024, 19206−19214



Wu Wenlong, born in 2001. Master candidate. Student member of CCF. His main research interests include large language model and knowledge graph.

吴文隆, 2001 年生. 硕士研究生. CCF 学生会员. 主要研究方向为大语言模型、知识图谱.



Yin Hailian, born in 1979. Associate professor, master supervisor. Her main research interests include multidisciplinary design optimization of aircraft and machine learning.

尹海莲, 1979 年生. 副研究员, 硕士生导师. 主要研究方向为飞行器多学科设计优化、机器学习.



Wang Ning, born in 1987. PhD, lecturer. His main research interests include mobile crowd sensing, database management systems, and wireless sensor networks.

王　宁, 1987 年生. 博士, 讲师. 主要研究方向为移动群智感知、数据库管理系统、无线传感器网络.



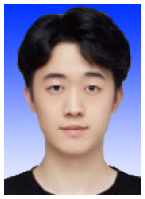
Xu Mengfei, born in 2000. Master candidate. Student member of CCF. His main research interests include knowledge graphs and large language models.

徐梦飞, 2000 年生. 硕士研究生. CCF 学生会员. 主要研究方向为知识图谱、大语言模型.



Zhao Xinzhe, born in 1998. Master candidate. Student member of CCF. His main research interests include large language models and knowledge graphs.

赵鑫喆, 1998 年生. 硕士研究生. CCF 学生会员. 主要研究方向为大语言模型、知识图谱.



Yin Zhanzuo, born in 2000. Master candidate. Student member of CCF. His main research interests include machine learning and multi-modal learning.

殷崭祚, 2000 年生. 硕士研究生. CCF 学生会员. 主要研究方向为机器学习、多模态学习.



Liu Yuanrui, born in 1997. Master candidate. Student member of CCF. His main research interest includes large language model.

刘元睿, 1997 年生. 硕士研究生. CCF 学生会员. 主要研究方向为大语言模型.



Wang Haofen, born in 1982. PhD, professor. Senior member of CCF. His main research interests include knowledge graph, natural lanugae processing, and large language model.

王昊奋, 1982 年生. 博士, 研究员. CCF 高级会员. 主要研究方向为知识图谱、自然语言处理、大语言模型.



Ding Yan, born in 1970. Master, senior-level engineer. Member of CCF. His main research interests include distributed databases and the field of big data.

丁　岩, 1970 年生. 硕士, 正高级工程师. CCF 会员. 主要研究方向为分布式数据库、大数据领域.



Li Bohan, born in 1979. PhD, associate professor. Distinguished member of CCF. His main research interests include spatiotemporal database, knowledge graph, large language model, and recommendation system.

李博涵, 1979 年生. 博士, 副教授. CCF 杰出会员. 主要研究方向为时空数据库、知识图谱、大语言模型、推荐系统.