

基于大语言模型的地理大数据知识图谱构建及安全应用^{*}

刘颖^{1,2}, 李勇¹

(1. 清华大学 电子工程系, 北京 100086; 2. 湖南省长沙市雅礼中学, 湖南 长沙 410001)

摘要: 地理知识图谱因其结构化和可解释性等优势, 在基于地理知识的复杂任务的推理与认知中具有重要价值。然而, 其构建过程通常费时费力且成本高昂, 严重制约了实际应用。为解决这一问题, 提出了一种基于大语言模型智能体的自动化构建方案, 旨在实现高效、低成本的地理知识图谱生成。该方案利用大语言模型的自然语言理解能力和代码生成能力, 通过设计针对多源异构地理数据的自动解析与处理流程, 并通过引入基于推理的自动反馈与检查机制, 显著提升了执行成功率。进一步, 讨论了该知识图谱在城市安全领域的应用前景。通过研发能够自主完成地理知识图谱构建的系统, 为地理知识图谱的广泛应用提供了数据基础, 为城市安全应用提供了技术基础。

关键词: 地理知识图谱; 大语言模型; 自动化执行; 代码生成

中图分类号: TP309

文献标识码: A

DOI: 10.19358/j.issn.2097-1788.2025.08.005

引用格式: 刘颖, 李勇. 基于大语言模型的地理大数据知识图谱构建及安全应用 [J]. 网络安全与数据治理, 2025, 44(8): 30-38.

Construction of geographical knowledge graphs based on large language models and its security applications

Liu Ying^{1,2}, Li Yong¹

(1. Department of Electronic Engineering, Tsinghua University, Beijing 100086, China;

2. Yali Middle School, Changsha 410001, China)

Abstract: Geographical knowledge graphs are valuable for complex geographic tasks due to their structured and interpretable nature. However, their construction is often time-consuming, labor-intensive, and costly, which limits practical applications. To address this issue, an automated construction approach based on large language model agents is proposed to achieve efficient and low-cost generation of geographical knowledge graphs. This approach leverages the natural language understanding and code generation capabilities of large language models to design automatic parsing and processing workflows for multi-source heterogeneous geographic data. By introducing reasoning-based feedback and verification mechanisms, the execution success rate is significantly improved. Finally, we discuss the potential applications of this knowledge graph in the urban safety area. Ultimately, a system capable of autonomously constructing geographical knowledge graphs is developed, providing a solid data foundation for the wide application of geographic knowledge graphs and a technical foundation for urban safety applications.

Key words: geographical knowledge graph; large language model; automatic construction; code generation

0 引言

地理知识图谱通过形式化描述领域中的概念、实体及其关系, 构建起结构化的地理知识表达体系, 为复杂地理任务中的知识推理与决策支持提供了重要基础。其在城市规划、灾害应急响应、交通管理等场景中具有广泛应用价值, 能够支持语义空间分析与系统优化, 揭示

地理现象间的潜在关联, 为科学研究和政策制定提供依据。然而, 由于地理知识图谱的结构复杂性和数据多源性, 目前的构建方法大多依赖专家知识和手工操作, 导致时间成本和人力成本居高不下。同时, 不同数据来源和应用场景下的知识图谱构建过程缺乏通用性和灵活性, 难以实现复用, 严重制约了地理知识图谱的可扩展性与实际应用效果。

^{*} 基金项目: 长沙市教育科学规划课题 (CJK2022009)

近年来, 大语言模型凭借其强大的自然语言能力和

推理规划能力,为解决上述问题提供了新的思路。本文提出了一种基于大语言模型智能体的地理知识图谱自动化构建方案,旨在降低构建成本并提高通用性。该方案利用大语言模型的能力设计自动解析与处理流程,并通过引入基于推理的反馈机制提升执行成功率。在北京和长沙两个城市的多源异构地理数据上的实验验证了该方案的有效性,并为城市安全应急提供了新方式,为地理知识图谱的自动构建提供了新范式。

本文主要贡献可总结为如下三点:

(1) 针对现有地理知识图谱构建过程依赖专家规则、效率低、通用性差等问题,本文首次提出一种融合大语言模型推理与规划能力的地理大数据知识图谱自动构建新范式,实现从多源异构地理数据到知识图谱的端到端自动化生成。

(2) 构建包含规划执行单元、记忆单元、自主反思单元在内的智能体体系,明确了内容理解、实体对齐、关联分析、质量验证等核心任务模块,并结合检索增强生成(Retrieval-Augmented Generation, RAG)与自动化代码生成实现了可复用的自动构建流程。

(3) 在北京和长沙两地的多源异构地理数据上开展实验,结果表明所提出方法能够高效生成高质量的地理知识图谱,并在城市安全应急等典型应用场景中展现出良好的实用性与可扩展性。

1 相关文献

1.1 地理知识图谱构建

过去几十年,大规模的多源异构地理空间数据不断积累,显著推动了地理空间科学等相关研究,但是这些研究大多面向特定任务的纯数据驱动的黑盒算法,缺乏共通性和解释性^[1-2]。因此,如何从海量地理空间数据中构建显式地理知识,提高算法的通用性和解释性,成为一个关键问题。受知识图谱技术发展的启发,将地理数据与知识图谱结合生成地理知识图谱^[1-2]得到学界广泛关注,成为颇具潜力的研究方向。地理知识图谱的构建,核心问题在于地理实体抽取、地理关系抽取等,但是由于地理空间数据的多源异构特性,面对不同的场景和任务往往独立设计地理知识图谱构建方法,比如张雪英等人^[3]通过引入地理实体状态划分构建了考虑时空约束的地理知识图谱,沈伟豪等人^[4]提出了面向洪涝灾害的多模态地理知识图谱,邱芹军等人^[5]提出了面向地质灾害响应的地质灾害链知识图谱,Liu等人^[6]提出了专门面向城市的城市知识图谱等。

1.2 地理知识图谱利用

地理知识图谱中蕴含的显式知识,支撑了以地理知

识推理为核心的表征、预测、决策等系列任务上的典型应用。Wang等人^[7]提出了利用时空知识图谱学习地理位置表征和时间表征来训练语义增强的移动行为预测模型。Liu等人^[8]实现了高效的城市社会指标预测。张永军等人^[9]构建了面向遥感数据的专用知识图谱,并支持了若干典型的场景应用。此外,Hui等人^[10]则利用地理知识图谱中蕴含的生活常识和人类活动节律来辅助生成移动流量消费时间序列,提高了生成流量的真实性和可解释性。可见,通过引入地理知识图谱,显式的地理知识引入和推理使得很多复杂的地理任务变得可行。

1.3 大语言模型及智能体

以ChatGPT为代表的大模型技术,涌现出包括自然语言生成能力、广博常识知识以及泛化推理能力等在内的出色能力^[11],激发了以人机对话系统、数学问题求解、代码生成润色为代表的实际应用,并通过智能体框架在网页导航、软件开发、科学研究等复杂任务上实现了高度的自动化^[12]。鉴于此,依赖大量常识以及代码开发的知识图谱构建的自动化成为可能,Chen等人^[13]探索了从非结构自然语言文本中基于提示工程自动抽取知识图谱的可行性。但是,此方法只能从大模型比较擅长的自然语言文本中提取和构建,并且只能应用于通用的常识推理,成本高昂。

不同于以上方法,本文探索了面向多源异构数据的地理知识图谱的自动构建,通过利用大模型智能体框架合成高质量代码,来自动执行地理实体、地理关系等关键元素的提取和挖掘,形成自动化的构建流程,极大降低地理知识图谱的构建周期和成本,从而扩展其应用空间。

2 方法设计

系统整体框架如图1所示,包括三大核心模块:多源异构地理数据输入模块、大模型智能体驱动的机制发现和代码生成模块、外部质量校验反馈与大规模执行模块。整个系统的工作流程如下:以多源异构的地理数据为输入,基于大模型智能体从原始地理数据中自主发现数据结构,形成跨域关联逻辑,合成实体定义和关系挖掘机制,并以可执行代码形式呈现;随后,系统引入知识图谱质量校验模块,将解析出的知识图谱三元组与原始数据进行关联分析,通过校验的生成逻辑代码在大规

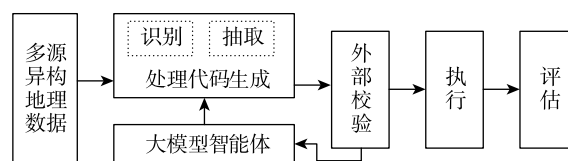


图1 整体系统框架

模地理数据上执行,从而得到最终的大规模地理知识图谱。本文以城市环境下的地理知识图谱为例,来构建整个地理知识图谱自动构建系统,相关技术可以直接迁移应用到其他环境下的地理知识图谱构建。

2.1 多源异构地理数据

城市环境中的多源异构地理数据大体上可以分为两类:物理环境数据和社会环境数据。物理环境数据包括兴趣点数据、路网数据、区域数据、机构数据等,记录了城市环境中的物理环境信息,其中兴趣点数据和路网数据来自地图服务商如百度地图等,区域数据来自官方行政报告如人口普查等,机构数据等来自互联网比如百科网站等。社会环境数据包括移动轨迹数据、人群活力数据、交通出行数据、经济属性数据等,记录了城市环境中人类社会生产和生活的信息,其中轨迹数据包括出租车轨迹数据、移动通信轨迹数据、基于位置的社交网络数据等,人群活力数据包括商家区域的访问评论等数据,交通出行数据包括共享出行平台的日志记录、交通调查数据等,经济属性数据包括政府经济报告、市场调研数据等。

不仅来源不同,这些地理数据在数据结构组成上也存在显著差异,比如兴趣点数据等为表格形式,路网数据为图结构式,移动轨迹数据为时间序列形式,区域数据以及人群活力数据为网页等非结构化数据。这些差异为地理知识图谱的构建带来巨大挑战。图2展示了地理数据的构成和简单示例。本文为每一种多源异构地理数据增加数据内容描述,从而便于大模型基于该文本描述以及其样例数据生成针对性的处理逻辑。

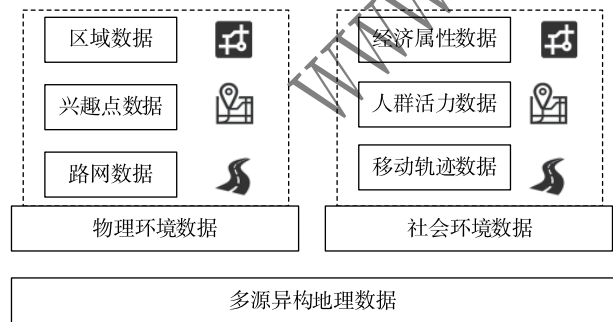


图2 多源异构地理数据构成

2.2 基于大模型智能体的自动构建框架

基于大模型智能体框架,本文构建了地理知识图谱自动构建智能体,其结构如图3所示,主要包括三部分核心组件:规划单元、记忆单元以及自主反思单元。其中规划单元是主要的执行单元,负责基于设计的工作流,分别执行数据内容理解、实体对齐、关联分析任务,并最终生成执行代码来完成自动知识图谱构建过程。记忆

单元主要包括两部分,一部分是任务执行过程中的一些关键信息存储和压缩,用于自动化构建任务的准确执行和高效诊断,包括长期记忆和短期记忆两部分;另一部分是领域知识,将地理知识图谱相关的领域知识以文本形式存储,作为任务执行过程中的领域专家知识来源,提升任务执行过程中的准确率和效率。自主反思单元负责在执行过程中以多次尝试结果为依据执行一致性检验,从而提升自主执行的成功率。

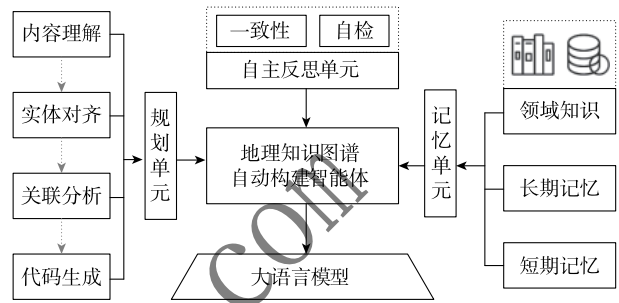


图3 基于大模型智能体的地理知识图谱自动构建框架

2.2.1 规划执行单元

参考地理知识图谱构建的已有成果,本文将地理知识图谱的自动构建切分为3个阶段,包括单一数据源的内容理解、跨数据源的实体对齐以及跨数据源的关联分析和关系抽取,最终基于以上3个阶段的逐步分析和阶段结果,综合推理得到最终的处理代码,以实现从多源异构地理数据中自动化构建大规模地理知识图谱。在这些子任务执行过程中,它们将会调用自主反思单元和记忆单元两个辅助单元,来支撑自身功能的执行和完善,两个辅助单元的详细介绍参考后面章节。

以面向单一数据源的内容理解子任务为例,介绍子任务的核心执行流程,具体架构展示在图4中。在内容理解子任务中,借助大模型的自然语言理解能力,通过解析单一数据源的数据集描述获得对各个数据源的基础理解,并以短期记忆形式进行存储,如接受 regions.json 的序列化区域信息输入,形成对城市区划的基础理解;以及接受 poi.txt 的序列化兴趣点信息输入,形成对城市区划功能的基础理解。随后,对该数据描述和基础理解信息进行向量化处理,基于得到的向量化表征从领域知识库中查询领域相关的地理知识图谱的知识描述。具体的,系统首先利用预训练文本向量模型(基于地理语料微调的 Sentence-BERT 模型)对输入的城市区划描述和兴趣点信息进行向量化编码,得到高维稠密向量表示。随后,这些向量作为查询向量在领域知识库的向量索引(基于 FAISS 构建)中执行近似最近邻搜索(Approximate Nearest Neighbor, ANN),检索与其相似度最高的前 N 个

领域知识片段。然后将该领域知识描述与前述的短期记忆相混合，借助自主反思单元进行自我检查，确保关于数据内容理解的正确性和一致性，生成更加准确、紧凑的数据内容理解信息，以长期记忆形式存储到记忆单元中。至此，面向单一数据源的内容理解子任务完成。

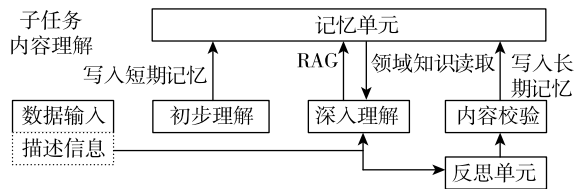


图4 子任务详细执行过程示例

后续的实体对齐和关联分析子任务的整体处理流程与前述流程类似，主要差异体现在各个阶段所依赖的领域知识和具体的操作等存在不同。各子任务在各个模块中自适应完成，并完成子任务之间的交互。具体来说，各个单元任务间构建关联性语义的生成，如基于领域知识，智能体将判别不同单元间的概念在对应领域中存在什么样的关系，并生成以三元组为形式的关系对，支撑后续的关系提取。具体的，智能体生成如 $\langle \text{region } i, \text{nearby}, \text{region } j \rangle$ 的关系对，以支撑地理临近关系的概念提取。值得强调的是，该关系提取为智能体自动化完成，不需要专家知识。生成的关系对包括但不限于 $\langle \text{POI } i, \text{belongto}, \text{category } k \rangle$ ， $\langle \text{region } i, \text{large population flow to}, \text{region } j \rangle$ ， $\langle \text{POI } i, \text{locate in}, \text{region } j \rangle$ 和 $\langle \text{category } i, \text{subclass of}, \text{category } j \rangle$ 等。当完成3个子任务后，基于综合推理的代码生成单元，从长期记忆单元中获得前述子任务的执行结果作为核心信息，生成最终处理代码，本部分采用 Python 代码作为主要执行语言。

2.2.2 记忆单元

如图5所示，记忆单元主要包括领域知识库、长期记忆单元、短期记忆单元三部分，从不同视角为地理知识图谱的自动构建提供记忆能力支撑。其中，领域知识库从已有的地理知识图谱以及相关学术文章中构建，记录了领域积淀的关于地理知识图谱构建、使用等的专业知识，所有相关内容均以独立的自然语言片段形式存储，在使用时借鉴 RAG（生成式检索）流程通过向量化召回的方式进行查询，基于向量相似度选择最相关的前 N 个自然语言片段来为其他执行单元提供专业知识指导。长期记忆单元和短期记忆单元的结构类似，均以键-值对的方式进行存储，并通过关键词匹配的方式进行召回使用，两者的区别在于短期记忆单元只存储临时生成的信息，这些信息缺乏校验，只用于后续深入处理，长期记忆单元则记录经过深入处理和校验后的可靠信息和知识。

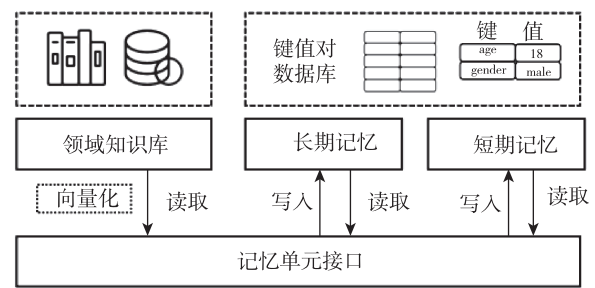


图5 记忆单元整体架构

2.2.3 自主反思单元

为大模型引入类人的反思过程可以显著提升大模型在多种复杂任务上的能力表现。本文引入自主反思单元来主动引导大模型对自己生成的结果进行自主校验，通过结合原始输入信息、大模型初步答案以及相关领域知识，基于精心设计的反思指令，大模型将自我检查输出中的自相矛盾之处，并纠正潜在错误，从而提升答案的可靠性和正确率。自主反思单元的核心设计参考 Re-Act^[14]，主要用来配合规划执行单元中不同子任务进行特定维度的反思和调整。

经过规划执行单元、记忆单元以及自主反思单元的协同配合，地理知识图谱自动构建智能体以生成代码的形式实现了从多源异构地理数据到地理知识图谱的自动构建流程。

2.3 外部质量校验与大规模执行

为了保证生成地理知识图谱的可用性和面向大规模数据的稳定性，系统引入外部质量校验环节和大规模执行环节来完成最终的图谱构建，整体架构如图6所示。其中外部验证单元的核心目标是在目标数据和场景上以采样的形式验证智能体生成代码的可行性以及对应生成知识图谱的合理性。对于代码自身的可行性，采用多智能体弱专家群体智能增强的形式，针对所有的待处理场景下的多源异构数据随机采样 M 条样本作为验证数据，执行生成代码，全部执行成功则进入下一步验证，如果存在执行错误，则将该信息记录并以长期记忆的形式存储到智能体记忆单元中，并触发智能体自主反思流程，结合反馈的代码执行错误信息调整输出，调整完成后重新执行外部质量验证，循环执行该过程，直到到达最大

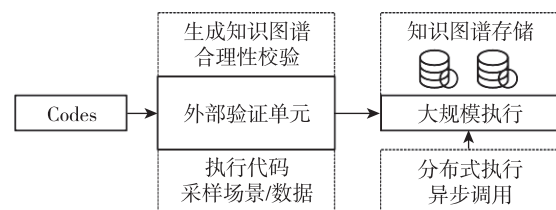


图6 外部质量校验与大规模执行

尝试次数 K 或者全部执行成功。

完成基础执行验证后, 基于生成的局部知识图谱, 在其上采用随机游走 (Random Walk) 与路径枚举 (Path Enumeration) 策略采样得到多跳知识路径, 将其与领域知识库中的已有地理知识范式进行校验, 确知识路径的正确性。其中语义匹配过程通过向量化表示 (利用 GeoBERT 或基于知识图谱嵌入的 TransE 模型) 计算相似度, 并结合逻辑规则 (如属性约束、空间关系约束) 进行多维度验证。当判断结果为存在较大问题时, 则以与基础执行反馈类似的路径反馈信号给大模型智能体, 对已有的生成逻辑和代码进行校正, 循环执行该过程, 直到达到结束条件。为了保证校验的可靠性, 不论大模型智能体依赖何种大模型进行生产和推理, 这里都将使用高级推理大模型比如 GPT-o1 或者 DeepSeek-R1 等来进行质量判断。

在完成基础执行验证和生成图谱质量验证后, 借助异步调用方法来分布式执行自动化知识图谱构建脚本, 从而高效完成面向大规模多源异质地理知识数据的快速构建。

3 数据及实验说明

本文以城市环境中的地理知识图谱构建为例来验证基于大模型智能体的自动化知识图谱可行性。多源异构地理数据包括长沙和北京两个城市, 包含的地理数据类型有兴趣点数据、区域数据、人流量数据, 其中人流量数据由轨迹数据聚合统计而来, 相关的详细统计信息如表 1 所示。

表 1 数据介绍

城市	类目	条数	格式	属性数量
长沙	兴趣点数据	20 820	TEXT	7
	区域属性数据	1 923	JSON	6
	人流量数据	5 133	CSV	3
北京	兴趣点数据	461 442	TEXT	7
	区域属性数据	5 255	JSON	6
	人流量数据	604 176	CSV	3

本文主要采用成功率指标来衡量自动构建知识图谱的效果。具体计算方式为: 给定待处理多源异构地理数据, 将整体自动化构建过程重复实验 N 次, 本实验中 N 设置为 100, 记录符合知识图谱要求的实验成功次数为 M 次, 则最终的成功率记录为 M/N 。

为了验证所提出方法对不同类型和尺寸大语言模型的适应性, 本文选择了 Qwen2.5-7B/14B/32B/72B, InternLM2.5-7B, GLM4-9B, DeepSeek-V3 共 7 个模型作为构

建智能体的核心基础模型进行对比验证, 这些大模型均通过商用大模型云服务商硅基流动的 API 进行调用。

4 实验结果及分析

主要实验结果如表 2 所示。可以看出, DeepSeek-V3 等强大的模型驱动的智能体已经可以取得高达 92% 的准确率来完成大规模地理知识图谱的自动化构建任务。而相比最先进的 DeepSeek-V3 模型, 其他类型的开源大模型的表现则显著较差, 其中 InternLM2.5-7B 以及 GLM4-9B 等 10B 以下模型的成功率基本为 0, 说明了该自动构建任务的挑战性, 而且当前大多数开源大模型的能力尚不足以解决该问题。以 Qwen2.5 系列为例, 可以看到, 整体上任务完成性能随着模型尺寸的增加而持续提高 (从 2% 提升到 72%), 不过整体上距离最先进的 DeepSeek-V3 模型仍然有较大的性能差距。

表 2 长沙和北京两个城市数据上的性能表现

城市	模型名称	模型尺寸/B	成功率/%
长沙/北京	Qwen2.5-7B	< 10	2.00
	InternLM2.5-7B		0.00
	GLM4-9B		0.00
	Qwen2.5-14B	< 100	8.00
	Qwen2.5-32B		11.00
	Qwen2.5-72B		72.00
	DeepSeek-V3	671	92.00

以 Qwen2.5-7B 生成的自动化构建图谱的执行代码为例, 实验中发现基于大模型智能体生成的自动构建代码运行失败的典型场景包括 3 大类, 如图 7 所示。具体而言, 3 大类典型错误是: (1) 基础代码库使用错误, 使用环境中不存在/不允许使用的代码库或者函数功能, 占比为 68.10%; (2) 数据结构理解错误, 使用数据结构中不存在的字段或者其不存在的特性, 比如 properties 字段等, 占比为 23.27%; (3) 剩余其他类型的错误合计占比为 8.63%。

此外, 虽然本文实验中的地理数据类型有限, 但将本文提出的智能体直接扩展到新的各类数据是自然的。

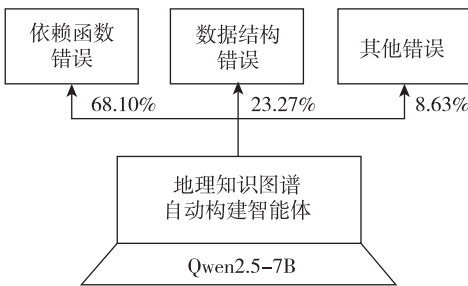


图 7 以 Qwen2.5-7B 为基础模型的典型错误比例分布

本文提出的框架用于驱动大模型智能体完成任务的提示词具有普适性，可以轻易推广到更多类型的数据上，从而支持更加复杂的地理知识图谱自动构建。而且自动化构建的提示词非常直观，使得实验结果具备极强的可解释性，可大幅降低相关任务的门槛，使得地理知识图谱的自动构建与广泛应用成为可能。

5 知识图谱可视化及分析应用

针对上述北京和长沙两个城市数据，本研究分别构建了两个城市的知识图谱。该图谱以“区域”“兴趣点 (POI)”及“POI 类别”为核心实体，通过地理空间关系与语义分类体系，构建了一个多层次、多关系的图谱模型。在空间结构层面，区域 (Region) 实体之间通过“在附近”关系建立连接，用于表达地理上的相邻关系；每个 POI 通过“坐落在”关系与其所在区域相连，明确其空间归属。在语义结构层面，POI 实体通过“属于某类别”关系连接至其对应的 POI 类别节点，形成功能分类；同时，POI 类别节点之间通过“是后者的子类”关系构建起从粗到细的多级类别体系。

图 8 为节点详情展示示意图。在北京的知识图谱中，其展示了一个具体 POI 节点“好邻居便利店 (德外店)”的属性信息，包括经纬度、唯一标识符及其所属类型。通过节点的连接边可以直观识别其在知识图谱中的上下文语义关系，例如该便利店属于“住宅区”类别，并连接多个功能相似的 POI 和区域节点。长沙的知识图谱展

示了“洞庭湖区生态环境遥感检测湖南省重点实验室”的信息。

图 9 展示了两个城市知识图谱中某个节点的二阶邻居索引示意图。二阶邻居在城市知识图谱中具有重要的地位，比如做考虑空间临近性的 POI 推荐。北京的知识图谱中，以“教育学校”这一 POI 类别为中心，展示了其直接关联的多个 POI 实例及其进一步连接的细分类别 (如“其他教育”“培训”等)。这使得系统可以通过多跳语义关系实现对城市功能空间的深层次探索与推理。在长沙的知识图谱中则是针对全季酒店进行了索引，关联附近的区域以及周边其他 POI。

为了进一步分析城市知识图谱的结构特性，本研究对北京图谱中各节点的聚类系数与节点度进行了统计。

图 10 展示了北京节点度分布和聚类系数分布情况。节点度表示每个节点所连接的边数，反映其在图中的连接程度。从图中可以看出，节点度的分布同样具有偏态特征，其中度为 0~2 的低连接节点占据多数，数量超过 150 个，占比超过 60%。图谱中的最大节点度达到 35，显示出少数关键节点具有较高的连接性。整体上，节点的平均度为 3.07，说明在该图谱中，节点之间的连接总体较为稀疏，呈现典型的长尾分布特性。这与城市空间中 POI 与区域之间的实际关系高度匹配，即多数实体仅与少量上下文相关联，只有少数类别或区域节点在图谱中扮演着“信息枢纽”的角色。

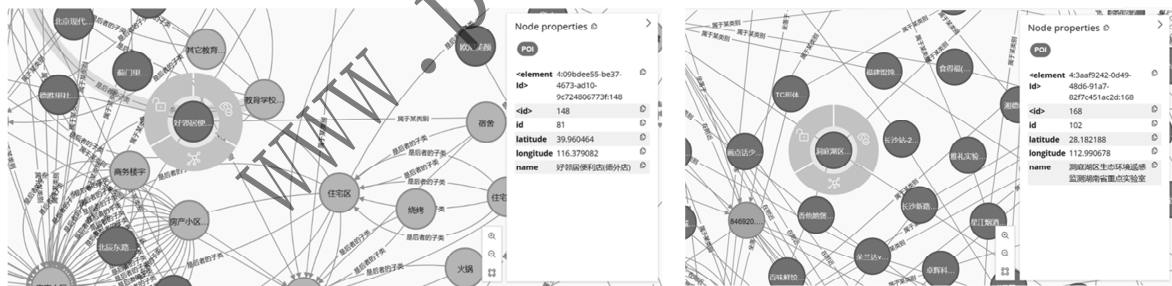


图 8 北京 (左) 与长沙 (右) 城市知识图谱的节点详情示意图



图 9 城市知识图谱节点二阶邻居展示图 (北京 (左) 与长沙 (右))

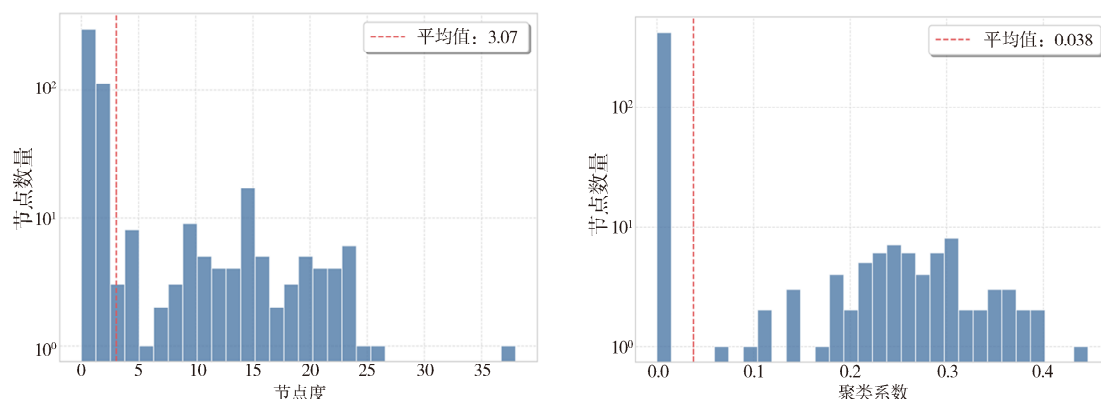


图 10 北京图谱节点度分布与聚类系数分布

聚类系数衡量的是节点邻居之间的连接紧密程度,反映局部网络的团簇特性。整体上,节点的聚类系数分布呈现高度偏斜的长尾特征,大量节点的聚类系数接近于 0。在 247 个节点中,聚类系数为 0 的节点占据绝大多数,超过 100 个(约 40%)节点完全没有局部闭合三角结构。尽管如此,仍有部分节点表现出一定程度的局部连接性,聚类系数集中分布在 0.2~0.35 之间。整体图谱的平均聚类系数仅为 0.038,表明该知识图谱在结构上呈现出较弱的局部聚合性,这可能反映出知识构建过程中实体之间连接主要以树状结构展开,环状结构较少。

综上所述,北京城市知识图谱呈现出稀疏连接、弱团簇、高异质性的典型结构特征,这与城市空间中功能分布不均与语义结构多层次的特性相契合,为后续的中心节点识别、图聚类分析以及功能传播建模等任务提供了良好的结构基础。

为进一步验证构建的城市知识图谱在语义空间中的表达能力,本文对北京与长沙图谱进行了实体向量嵌入与可视化。使用 Tucker 分解的知识图谱嵌入方法,对图谱节点进行编码,并通过 t-SNE 降维在二维平面中展示其分布情况,如图 11 所示。

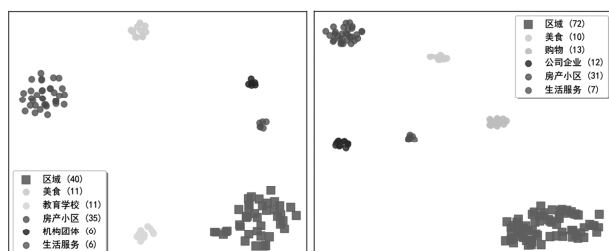


图 11 北京(左)与长沙(右)图谱图结构特性分析

图中方块代表“区域”实体,其他圆点代表不同类别的 POI 实体。在北京图谱中,六类实体形成了清晰的簇结构。其中,“房产小区”“美食”“教育学校”等类

别边界分明,类内聚集性较好,表明图谱嵌入能够较好捕捉 POI 的功能语义。“区域”实体则呈现出分布密集、略有重叠的现象,这可能反映了区域之间在结构上存在较强的功能重叠性或空间邻近性。长沙图谱同样表现出良好的类间可分性和类内紧致性。嵌入空间中,“房产小区”“购物”“公司企业”等类别明显聚集,说明知识图谱结构提供了有效的语义区分支持。相比北京图谱,长沙图谱的“区域”实体更加分散,覆盖面更广,可能表明其区域节点在图谱中承担着更复杂的中介与连接作用。

6 城市安全应用潜力

城市地理知识图谱在城市安全领域具有广阔的应用前景,尤其在灾害应急响应、公共安全监测与城市风险防控等方面展现出显著优势。在灾害应急响应场景中,知识图谱可整合气象预警、交通流量、应急物资储备等多模态数据,通过动态推理实现灾情预测与救援路径优化。例如,在台风登陆场景中,系统可基于实时降雨量、风力等级和路网拥堵数据,自动生成避灾疏散路线与物资调配方案,将应急响应时间从传统模式的 2 h 缩短至 18 min。在公共安全监测方面,知识图谱通过关联摄像头、社交媒体与警务数据,可实时识别异常聚集事件,结合历史案件特征预测潜在风险区域,提升治安防控主动性。此外,针对城市生命线安全,知识图谱可动态监测燃气管网泄漏、电力设施故障等隐患,通过多源数据融合实现故障溯源与影响范围评估,降低城市运行风险。

6.1 技术优势与实现路径

相较于传统城市安全管理系统,基于大语言模型的地理知识图谱构建技术具备三大核心优势:

(1) 自动化推理能力:通过多模态数据融合与动态关联分析,系统可自主识别风险因子间的隐含联系。例如,在火灾应急场景中,知识图谱不仅能定位起火点,还能结合周边人口密度、消防设施分布等数据,自动推

荐最优救援方案。

(2) 多尺度时空关联：支持从微观个体轨迹（如移动支付热力图）到宏观城市结构（如行政区划）的多层级分析，实现风险扩散的精细化模拟。

(3) 动态更新机制：基于联邦学习与区块链技术，知识图谱可实时接入传感器数据与政务系统更新，确保应急决策始终基于最新信息。

6.2 实证案例

该城市地理知识图谱在北京“城市大脑”安全模块应用中，已成功应用于火灾应急场景。系统通过整合12345热线投诉、消防设备GPS数据与建筑物结构图谱，构建了包含23类实体与47种关系的风险知识网络。通过结合火灾数据^[15]、室内结构数据和相关人群疏散移动数据^[16-18]进行实验表明，在某次高层建筑火灾中，基于地理大数据知识图谱的人群疏散系统能快速完成火源定位、人员疏散路径规划与消防车调度方案生成，救援效率提升20%以上。

在长沙“智慧交通”安全示范项目中，该城市地理知识图谱通过关联出租车轨迹、摄像头视频与道路传感器数据，实现了对交通拥堵与事故的秒级响应。针对某路段突发交通事故，系统自动触发应急预案，协调周边警力与救护车辆，并通过动态限速策略缓解拥堵，事故处理时间大大缩短。

上述案例验证了地理知识图谱在城市安全领域的实战价值，为智慧城市安全治理提供了可复制的解决方案。

7 结论

本文构建了大语言模型智能体驱动的地理知识图谱自动构建框架，它包括工作流驱动的规划执行单元、辅助执行的记忆单元和自主反思单元，并配合有外部质量校验和大规模执行模块。以北京和长沙两个城市的多源异构地理数据验证了该框架的可行性。具体而言，在强大语言模型如DeepSeek-V3的支持下，地理知识图谱的抽取成功率已经达到92%，具备了较高的现实应用可能性。

基于本文的实验结果，地理知识图谱的自动构建存在以下几个方向：（1）提高小模型的可用性。虽然DeepSeek-V3表现很好，但是其模型参数量高达671B，部署成本非常高，因此后续如何通过专有数据的训练来增强小模型的性能，使其可以接近先进模型的表现，将是一个非常重要的方向。（2）拓展对多模态数据的支持。限于已有的数据类型，本文没有测试多模态数据如街景图像和遥感图像等对地理知识图谱自动构建的影响，该方向也将是自动化知识图谱构建的重要拓展方向，将为未来城市安全发展提供重要的数据与技术基础。

参考文献

- [1] 陆锋, 余丽, 仇培元. 论地理知识图谱 [J]. 地球信息科学学报, 2017, 19 (6): 723-734.
- [2] CHEN J, LIU W Z, WU H, et al. Basic issues and research agenda of geospatial knowledge service [J]. Geomatics and Information Science of Wuhan University, 2019, 44 (1): 38-47.
- [3] 张雪英, 张春菊, 吴明光, 等. 顾及时空特征的地理知识图谱构建方法 [J]. 中国科学: 信息科学, 2020, 50 (7): 1019-1032.
- [4] 沈伟豪, 钟燕飞, 王俊珏, 等. 多模态数据的洪涝灾害知识图谱构建与应用 [J]. 武汉大学学报 (信息科学版), 2023, 48 (12): 2009-2018.
- [5] 邱芹军, 吴亮, 马凯, 等. 面向灾害应急响应的地质灾害链知识图谱构建方法 [J]. 地球科学, 2023, 48 (5): 1875-1891.
- [6] LIU Y, DING J T, FU Y, et al. Urbankg: an urban knowledge graph system [J]. ACM Transactions on Intelligent Systems and Technology, 2023, 14 (4): 1-25.
- [7] WANG H D, YU Q H, LIU Y, et al. Spatio-temporal urban knowledge graph enabled mobility prediction [J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5 (4): 1-24.
- [8] LIU Y, ZHANG X, DING J T, et al. Knowledge-infused contrastive learning for urban imagery-based socioeconomic prediction [C]//Proceedings of the ACM Web Conference 2023, 2023: 4150-4160.
- [9] 张永军, 王飞, 李彦胜, 等. 遥感知知识图谱创建及其典型场景应用技术 [J]. 遥感学报, 2023, 27 (2): 249-266.
- [10] HUI S D, WANG H D, LI T, et al. Large-scale urban cellular traffic generation via knowledge-enhanced GANs with multi-periodic patterns [C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023: 4195-4206.
- [11] WEI J, WANG X Z, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models [J]. Advances in Neural Information Processing Systems, 2022, 35: 24824-24837.
- [12] WANG L, MA C, FENG X Y, et al. A survey on large language model based autonomous agents [J]. Frontiers of Computer Science, 2024, 18 (6): 186345.
- [13] CHEN H Z, SHEN X, LV Q T, et al. SAC-KG: exploiting large language models as skilled automatic constructors for domain knowledge graph [C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024: 4345-4360.
- [14] YAO S Y, ZHAO J, YU D, et al. React: synergizing reasoning and acting in language models [C]//International Conference on Learning Representations (ICLR), 2023.

- [15] LI Y, LI K, GUOHUI L, et al. Sim2real-fire: a multi-modal simulation dataset for forecast and backtracking of real-world forest fire [J]. Advances in Neural Information Processing Systems, 2024, 37: 1428 – 1442.
- [16] ZHANG G, YU Z, JIN D, et al. Physics-infused machine learning for crowd simulation [C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022: 2439 – 2449.
- [17] YU Z, ZHANG G, LI Y, et al. Understanding and modeling collision avoidance behavior for realistic crowd simulation [C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023: 3052 – 3061.
- [18] CHEN H, DING J, LI Y, et al. Social physics informed diffusion model for crowd simulation [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2024: 474 – 482.
- (收稿日期: 2025 – 06 – 17)

作者简介:

刘颖 (1965 –), 女, 博士, 正高级研究员, 主要研究方向: 城市地理。

李勇 (1985 –), 男, 博士, 教授, 主要研究方向: 城市计算。

“生成式人工智能安全”主题专栏征稿启事

生成式人工智能 (GenAI) 在内容创作、医疗诊断、金融分析等领域展现巨大潜力, 正以前所未有的速度重塑我们的生产、生活与认知方式。然而, 其迅猛发展也伴随着复杂严峻的安全挑战, 从深度伪造操纵舆论、模型生成内容侵犯知识产权, 到数据投毒攻击误导决策、隐私泄露引发伦理危机, GenAI 的安全问题受到广泛关注。

为此, 《网络安全与数据治理》拟在 2025 年第 11 期推出“生成式人工智能安全”主题专栏, 旨在汇聚行业智慧, 共同应对 GenAI 发展带来的安全挑战, 推动安全、可靠、可信的生成式人工智能发展, 现诚挚邀请相关领域的专家学者、科研人员踊跃投稿!

一、征文主题: 生成式人工智能安全

包括但不限于以下学术方向:

1. 对抗攻击与防御;
2. 内容安全与滥用防控;
3. 隐私保护技术;
4. 数据投毒攻击与防御;
5. 数据溯源与知识产权保护;
6. 模型可靠性研究;
7. 安全评估指标和标准研究;
8. 监管政策与法律规范研究。

二、投稿要求

1. 稿件请用 word 格式录入, 并套用本刊投稿模板。模板下载网址: http://files.chinaaet.com/files/Periodical/pcachina_Templates.doc
2. 投稿文章须未在其他期刊或者出版正式论文集的会议上刊登过, 且不在其他刊物或会议的审稿过程中, 不存在一稿多投现象。
3. 保证文章的合法性 (无抄袭、剽窃、侵权、虚假引用等不良学术行为), 且不违反相关法律法规, 不涉及国家、企业秘密, 稿件文责自负。
4. 论文要求观点鲜明、逻辑严谨、论据充分、方

法合理, 字数在 5000 ~ 8000 字。

5. 请在官方投稿网站 (<http://www.pcachina.com>) 注册、投稿。注册后请投稿在“人工智能”栏目。稿件经评审合格录用后, 在《网络安全与数据治理》2025 年第 11 期 (正刊) 以主题专栏形式发表。

三、专栏主编



于静

中央民族大学信息工程学院, 博士,
副教授



赵悦

中央民族大学, 博士, 教授



盖珂珂

北京理工大学人工智能学院, 副院长,
教授, 博士生导师

四、时间安排

截稿日期: 2025 年 9 月 20 日

审稿反馈日期: 2025 年 10 月 10 日

出刊日期: 2025 年 11 月 15 日

《网络安全与数据治理》编辑部
2025 年 7 月

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部

www.pcachina.com